AN OPEN SOURCE APPROACH TO SOCIAL MEDIA DATA GATHERING

THESIS

Anthony J. Kallhoff, 2nd Lieutenant, USAF

AFIT-ENS-MS-18-M-130

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-130

AN OPEN SOURCE APPROACH TO SOCIAL MEDIA DATA GATHERING

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Anthony J. Kallhoff, BS

2nd Lieutenant, USAF

March 2018

AFIT-ENS-MS-18-M-130

AN OPEN SOURCE APPROACH TO SOCIAL MEDIA DATA GATHERING

Anthony J. Kallhoff, BS

2nd Lieutenant, USAF

Committee Membership:

Dr. Bradley Boehmke
Chair

Maj. Jason Freels
Reader

AFIT-ENS-MS-18-M-130

Abstract

Modern usage of social media affords the military intelligence and analytic communities novel approaches to gather information. However, the tools and resources to develop these methodologies are still maturing. Furthermore, current data acquisition tools are not available to the DoD for all social media platforms. This thesis addresses a small subset of this problem by developing an open source methodological approach to collect and manage data from a popular social media site that has previously been inaccessible to defense intelligence organizations. This approach was operationalized via the R package called `instaExtract,` and an exemplar analysis was performed to demonstrate its application and efficiency for intelligence gathering.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Bradley Boehmke, for his guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would also like to thank my sponsor for both the support and latitude provided to me in this endeavor.

Anthony J. Kallhoff

Table of Contents

Page

List of Figures

Page

List of Tables

Page

AN OPEN SOURCE APPROACH TO SOCIAL MEDIA DATA GATHERING

I. Introduction

1.1 General Issue

The modern world is producing mountains of information that can be of indispensable benefit to the Department of Defense (DoD) intelligence community. Traditionally, intelligence has been gathered through arduous means and required advanced systems or highly trained professionals. However, the emergence of social media, and the breadth of new information it provides, offer analysts a unique opportunity to access large amounts of data at little cost. While the techniques to make the most of this data are still in development, they present an exciting supplement to traditional intelligence gathering.

In academic and applied research, there is an emphasis on advancing and applying rigorous analytical techniques to data, but also important are the methods to acquire the data. Just as higher resolution microscopes provide new insights in biology, tools that can offer new and refined streams of information can help to better inform analysts. Without attention being given to the acquisition of data, analysts are forced to apply their techniques to flawed or constrained data sets. Current data gathering tools have varying limitation in their scope, functionality, and platform, restricting the benefits they may offer.

1.2 Research Goals

The current suite of in-house data acquisition tools does not span all the various social media platforms. One notable site that is presently excluded is Instagram, a photo and video sharing platform. While there exist third-party programs that offer some form of information gathering from Instagram, they are limited in varying ways. This research aims to establish a reliable and reproducible approach to extract Instagram data for use by the DoD intelligence community. To this end, this research seeks to create an R-based Instagram data acquisition package with the following functionality:

- Information retrieval without the need for user authorization or API integration
- Search functions for identifying users and hashtags in Instagram's databases
- Media post retrieval that includes meta-data and information on community involvement
- Retrieval of information for users and their activity on the site
- Raw data stream cleaning and transformation resulting in user friendly data
- Example analytic functions to serve as a basis for future development
- Hosting on a platform that facilitates community involvement, continued support, and version control

1.3 Research Contributions

This research aims to expand the DoD intelligence community's information gathering capabilities with a consistent and robust approach to pulling Instagram data. This approach will be in the form of an R package hosted on the Air Force Institute of Technology (AFIT) Data Science Lab's GitHub account, consistent with the approach of recent analytic tools developed by AFIT. More than a one-off deliverable, the results of

this work will establish a groundwork for more Instagram based intelligence gathering and analysis. By focusing on the particulars of data acquisition, this research supports the efforts of data analytics by providing the material to investigate. This package will also have an immediate impact on the DoD, who will have access to the information and insights Instagram provides.

1.4 Assumptions and Limitations

In development of these resources, some level of continued success by Instagram and the usefulness of its information is assumed. The tools developed also rely on a general structure to how Instagram presents its data and allows users to interact with that data. Significant reworkings to Instagram's procedures might invalidate some functions in this research.

1.5 Organization

The paper begins with a literature review in Chapter 2 that explores intelligence gathering and its evolution with social media. Furthermore, Chapter 2 looks at the current resources for collection social media data, diving specifically into Instagram, its layout and tools that interact with it. Finally, Chapter 2 addresses the topic of reproducibility in analysis. Chapter 3 discusses the methodological framework for extracting Instagram social media data and the open source R package, `instaExtract`, that operationalizes this approach. Next, Chapter 4 presents an exemplar analysis, showing how an analysis might utilize `instaExtract`. Finally, Chapter 5 summarizes the work accomplished with this thesis and discusses areas for future research.

II. Literature Review

2.1 Overview

The objective of this thesis is to develop and operationalize a methodological approach to collect and manage data from a popular social media site as well as provide an R package that allows for reproducible analysis by users. To this end, this paper begins with a literature review broken into six sections, including the overview. The second section will be a look at the more traditional fields of intelligence gathering and how computers are being integrated with older methods. The third section will look closer at the specific social media platform explored in this thesis, Instagram, and the structure of its information. The fourth section covers practices for collecting and managing large amounts of user data pulled from a social media site using its application program interface (API). Fifth, the review examines existing data extraction software and their advantages and limitations. The final section explores techniques and considerations to ensure reproducible analysis.

2.2 The Evolving State of Intelligence Gathering

The gathering of intelligence has developed throughout the history of warfare. Knowledge of the enemy and of the landscape of the battle can be vital to the success of the combatants. Accordingly, the U.S. has developed many institutions and practices for the purpose of acquiring and deciphering intelligence (Richelson, 2015). Many of the techniques relied on, however, are fallible and inconsistent. Combing through traditional paperwork can take many man hours, and even experienced eyes can miss important details. Many human intelligence procedures also depend on investigators trusting their memories or interpreting ambiguous signs (Loftus, 2011). With the advent of computers,

the nature of intelligence gathering is shifting rapidly. Digitized files can be searched in milliseconds, and while computers carry their own sets of faults, they do not tire or relent.

More than just work horses though, computers offer unique opportunities for data acquisition and analysis (Leese, 2015). With the internet came the introduction and spread of social media sites like Facebook and Instagram. These sites are populated with information from users all over the world and represent almost all demographics. There is such a wealth of information that has sprung up so recently, that the intelligence community is struggling to adapt fast enough to utilize all the data now at its disposal (Loftus, 2011). Along with a flood of new information, computers are powerful processers of this information. As an example, consider the analysis of terrorist networks. Developing an organizational structure of terrorist's cells has been a backbone of human intelligence in the fight against terror. Isolating important targets and determining the flow of information has allowed us to neutralize cells with minimal damage (Ronczkowski, 2011). However, these models are currently still heavily reliant on the same unreliable techniques mentioned early. Computers and the new sources of information they provide could afford the intelligence communities new manners of network analysis and formulation. Specifically, social media can provide information unlike any other source. Social media data provides relational and geospatial data at a real time pace, closing the turnaround of reports for analysts (Pitic, Volovici, Tara, & Mite, 2013).

The structure of social media sites and the way users interact with them provide new types of information to be used in analysis. One of the primary and most useful of these types comes in the form of social networks, a network of actor nodes and

relationship or interaction edges that bind them together (Aggarwal, 2011). While social networks have been studied in conventional scenarios for some time, social media data presents itself in a manner so conducive to social network analysis that it has created a renewed fervor in the scientific study of the field (Aggarwal, 2011). These online social networks tend to reach large scales which manifest with special structures and behaviors (Milgram, 1967). Online networks in particular are saturated with data, offering unique circumstances for analysis of two primary kinds of data: structural analysis which identifies important nodes, links and regions, and content-based analysis which delves into the specifics of the content being shared (Aggarwal, 2011). Finally, online networks evolve in much shorter time spans than conventional networks, a complication that is the subject of recent research (Aggarwal & Yu, 2005).

The previous paragraph mentions content-based analysis briefly, but the scope of analysis to be done on social media data is extensive. Almost all social media platforms are rich in text based information which lends itself to interesting fields of analysis like classification, clustering, and sentiment analysis (Aggarwal, 2011; Ashraf, Verma, & Tech, 2016). Often, social media data will also contain image and video content, enabling methods involving image and voice recognition (Li, Zha, Huet, & Tian, 2016). One other category of information, geotagging, is a descriptor of content, but can provide for meaningful analysis in its own right. Location information can help to contextualize online content in the real world (Andriopoulou & Lymberopoulos, 2012).

2.3 Instagram

Instagram launched on 6 October 2010 with 25,000 sign-ups on the first day ("Our Story – Instagram," n.d.). Two years later, on 9 April 2012, the company, now at

30 million users, was purchased by Facebook for nearly $1 billion (Chaykowski, 2016). Today, it boasts 500 million daily users, and more than 800 million monthly active users ("Our Story – Instagram," n.d.). Instagram is, at its core, just a photo-sharing app, but with more users than Twitter, Snapchat, and Pinterest combined, it has placed itself as one of the fastest growing social media platforms of all time (Chaykowski, 2016).

The typical Instagram experience revolves around two types of content: a post and a story. A post, seen in Figure 4, is the term for a user submission encompassing their submitted media, description, and community response. A post can either be a video or a photo. It is accompanied with a user created caption and hashtags, markers that promote and categorize a post. Posts can also have comments and likes left by other users. Once a post is created, it will exist on a user's account indefinitely unless it is deleted. A story is similar to a post, but it is only up for a short time and comments and likes are not displayed. Instead, stories act as a slideshow that move through videos and photos posted by the accounts a user follows. Others can comment on a story, but the comment and who viewed a story is only visible to the creator. Currently, Instagram only supports mobile users viewing the story of accounts that they do not follow.

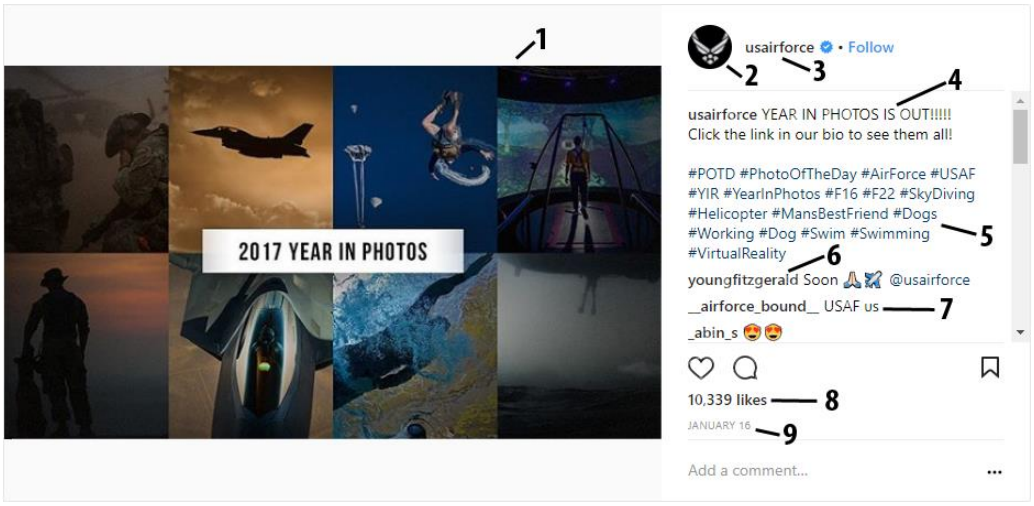

Figure 1. An Instagram Post: 1) Media – can be a video or photo, 2) Owner Thumbnail, 3) Owner Username, 4) Post Caption, 5) Post Tags, 6) Commenter Username, 7) Comment, 8) Post Likes, 9) Post Date

A user's history of posts is stored as an account, shown in Figure 2. An account page shows the user's thumbnail and username. It also has statistics on how many posts the account has made, how many other users follow the account, and how many others the user follows. To follow an account means to have the posts of that account appear in the stream of posts provided to the user, called their feed. If someone is signed into Instagram, it also possible to view which accounts are followers and which accounts the given account is following. Next, accounts have a bio that may describe their content or themselves. Finally, on the account page are listed the posts of the user from most recent to oldest.

Figure 2. An Instagram Account

Similar to an account page, Instagram has pages for locations. Locations are created by users, and subsequently used by others to classify future posts. An example location is shown in Figure 3. The page shows the coordinates of location, a story that users can add content to, the current top nine posts, and the list of all other content in reverse chronological order. Since locations are user created, there is not an extensive list of all locations, but it is possible to view the top 1000 locations for a city by navigating the options of the "explore/locations" subdomain.

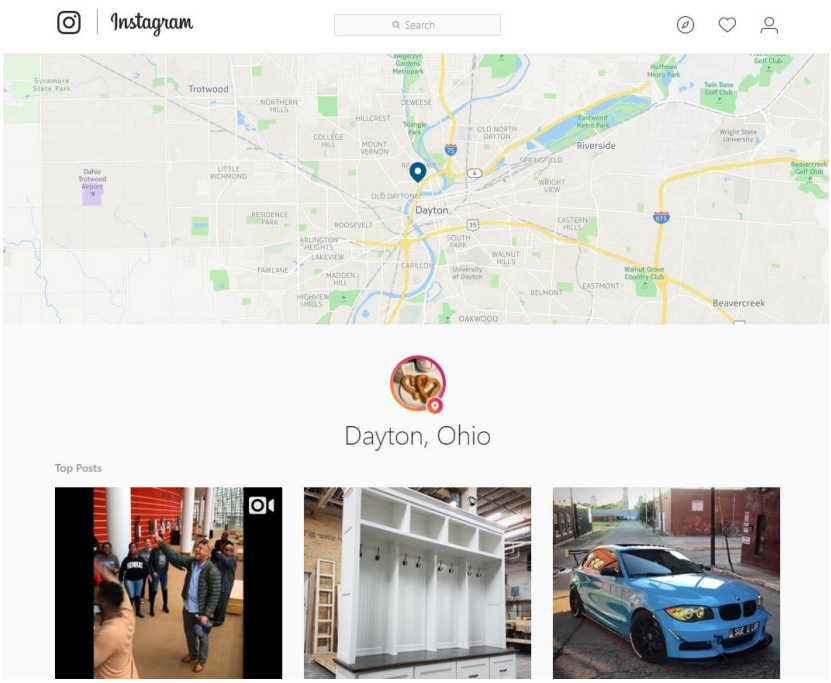

Figure 3. An Instagram Location

2.4 Data Collection and Management

Data analysis can be thought of as a continuing process rather than a one-step application. This analytic cycle is depicted in Figure 4. Although much of the work in the process is done in the sub-section labeled “understand”, it is also important to not neglect the first two steps, “import” and “tidy.” Importing entails pulling information from a source, whether it is already organized or through more intensive processes like web scraping. Tidying information is the process of formatting data into a usable and consistent structure (Wickham & Grolemund, 2016). These two steps are essential to the analytic process since without them, the analyst has no data to work with or the data exists in a form that cannot be analyzed. Therefore, in the pursuit of analyzing social media information, attention should be given to the importing and tidying of the data, as is the emphasis of this thesis.

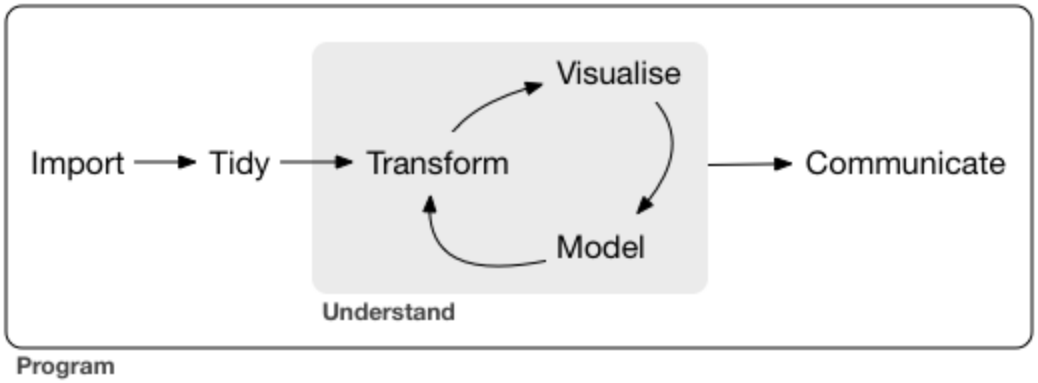

Figure 4. The Analytic Cycle (Wickham & Grolemund, 2016)

The first step to acquiring information from a social media site is determining an avenue to the information. One source of information is to scrape the desired fields directly from the displayed page, commonly referred to as web-scraping. Such a program would traverse the actual webpage that is displayed when a person uses the site. To move between pages, it would use built in links or travel along links in the page to search for requested information. This web-scraping approach relies on the underlying structure of HTML-based sites to process what was made for human interaction (Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato, & Fdez-Riverola, 2013). This approach offers complications to a program since it does not adapt well to changes to the site and requires specific implementation for each portion of a site (Glez-Peña et al., 2013). Most websites also offer an application programming interface (API) which allows outside organizations access to the information stored in the sites databases. APIs offer a more consistent means of retrieving data in consistent formats, but it is not guaranteed that they will offer all information that the site stores (Glez-Peña et al., 2013). Additionally, as is the case with Instagram, APIs may be subject to strict regulations that prevent the public from accessing them. The alternative to this is an open API, which allows third-party

developers access without needing authorization, but it can still suffer the other problems of APIs (Jang & Lee, 2009). With both APIs and web scraping, sites can be behind an authentication wall that will not allow some or all of their information to be displayed unless the user is logged into an account.

Once a method of access to the website is determined, the information is still in need of storage and management. The information collected is still in a raw state and as a stream is difficult to work with (Injadat, Salo, & Nassif, 2016). When web scraping, the data can be stored in almost any format, depending on how the web scraping software pulls its information. APIs will return a data structure based on established guidelines, one of the benefits of interacting through an API (Glez-Peña et al., 2013). In the case of Instagram's open API, data is return in JavaScript Object Notation (JSON). JSON files are fundamentally text files, but they are organized in a way to make them both readable for humans and easy for machines to parse ("JSON," n.d.). A complete list of data formats would be infeasible, but the basic principle is that data will come in a certain format depending on the chosen importing method.

Whatever the method, data will be required to conform to certain requirements in order to be compatible with the needs of tools and data storage later in the process. This conversion of data falls under the category of "tidying" the data, or ensuring that the data is in format that best facilitates future work(Wickham & Grolemund, 2016). Toward this goal, the programming language R has resources that facilitate the "tidying" process. R is an open-source programming language with a focus on statistical computing and extensibility ("R: What is R?," n.d.). R is extended via packages, fundamental units of reproducible code (Hadley Wickham, 2015). One such package, `tidyr`, specializes in

tidying data (Hadley & Henry, 2018). This package and more like it make R a great programming language for manipulating data of many different formats and preparing it for use in analysis.

A question not yet answered is what information is important to pull? It is impractical to have access to all the data flowing through social media sites without the cooperation of the company behind them, especially for sites as large as Instagram. A decision must be made then to choose what data to target and how to find the data relating to that target by second order connections. While the ultimate use of information and methods of analyzing are outside the scope of this thesis, exploring potential use cases allows for the creation of a program that more accurately reflects the needs of future users.

One strategy for analysis is a user focused approach. By viewing the post rate and other statistics like following or resubmissions, analysts can put together a map of influential users and their reach (Erlandsson, Bródka, Borg, & Johnson, 2016). Further, they can track the flow of information and events to construct communities and networks, similar to those created now with more conventional gathering techniques (Lv & Guo, 2016). Networks can also be established to connect users. Algorithms that watch the behavior of users can flag behaviors and interests that share a high correlation with target user bases, providing new insight into potential persons and groups of interest (Mezghani, Péninou, Zayani, Amous, & Sèdes, 2017). This is an example of a unique application of computers, where it may discover underlying connections that even an experienced investigator might miss.

Another angle for analysis is to focus on events rather than networks. Social media platforms have already shown the potential to be important barometers in public perception of events and topics. By following trends, hashtags, and locations of posts, analysists can create models that capture the sentiment and movement of a target audience (Bian et al., 2016). This leads again to an area where computer driven analytics can greatly help in information gathering. In a world of social media and interconnectedness, discussion of events can evolve faster than a system administrator can have time to adapt to. So it becomes important for the information gathering software to have ways of dynamically associating common terms and pulling from sources that a conventional search might overlook (Chan, Vasardani, & Winter, 2014).

2.5 Existing Instagram Data Extraction Software

Tools for extracting information from Instagram currently exist but are limited in their usefulness. An overview of some available solutions is presented in Table 1. This review will look more closely at `instaR` and `Instagram PHP Scraper`, but looking at Table 1 shows that the various programs approach the same problem in many ways. There are free programs like `Instagram-scraper` by Richard Arcega (Arcega, 2018) and monetized software like the Instagram scraper provided by Im Rista ("Instagram Scraper - Im Risto - Internet Marketing Blog," n.d.). Many of the programs are based on open source languages like Python and PHP, but only `instaR` notably uses R. The monetized programs do not make their source code available, and as such, would not easily be extendable. The existing software serves as a framework to inform functionality, but ultimately each program fails to capture every important feature. The monetized software has restrictive licensing issues and closed code. Programs in PHP and

Python do not capitalize on the existing analytic tools employed by the DoD community. And lastly, `instaR` relies on interactions with Instagram's private API, and the authorization required to use that do not fit with mission needs.

Table 1 - Instagram Data Extraction Software: O – Optional, G – GitHub, ? – Unavailable Information

	instagram-scraper	Instagram Super Scraper	Instagram PHP Scraper	Im Risto	InstaScraper	instaR
R						✓
Python	✓					
PHP		✓	✓			
Custom				✓	✓	
API				✓	?	✓
Scrapping	✓				?	
JSON	✓	✓	✓		?	
Authentication	O	O	O	✓	?	
Authorization						✓
Availability	G	G	G	$49	$8	G
GUI		✓		✓	✓	

These programs, however, can still be useful as templates for this research. First, this thesis explores the `instaR` package by Pablo Barberá (Barberá, 2017). The program is an R-based package that allows the user to interact with Instagram's API to perform a range of functions that are detailed in Table 2. The functionality of `instaR` is not as expansive as other programs, but it highlights a few important considerations for Instagram scrapers. Besides being outside of DoD control, `instaR` would have been a good choice for Instagram data extraction based in R if it didn't require authorization.

Since November of 2015, all API endpoints require a valid access_token, which requires the application to be reviewed and approved by Instagram themselves ("Platform Changelog • Instagram Developer Documentation," n.d.). This is already a difficult process for civilian users, but is a severe deterrent for the use of the API in a DoD capacity. However, working with the API has two advantages Instagram-php-scraper, the program investigated next, does not have. First, working with the API generally ensures a higher level of stability with how a function will behave. Secondly, the API has some functions that are hard to replicate through other means, namely, the ability to search for posts in a given region.

Table 2 - instaR Functions

Function	Purpose
getComments	Retrieves up to 150 recent comments for a given post
getFollowers	Retrieves the list of users that follow a given user
getFollows	Retrieves the list of users a given user follows
getLikes	Retrieves the list of users who liked a post
getLocation	Retrieves location information
getPopular	Retrieves up to 24 popular posts
getTagCount	Retrieves a count of times a hashtag has been used
getUser	Retrieves public information about a user
getUserMedia	Retrieves public media from a given user and can download media
instaOAuth	Creates an OAuth access token enabling R to communicate with Instagram's API
searchInstagram	Search media for mention of hashtag or for posts in a given location

The other program investigated in detail is `Instagram-php-scraper` by Postaddict.me (Postaddict.me, 2018), whose functions are detailed in Table 3. Unlike

Table 3 - Instagram-php-scraper Functions

Function	Purpose
withCredentials	Allows the program to run with account information
searchTagsByTagName	Returns the hashtag results in a search for a given tag
getErrorBody	An error handling function
searchAccountsByUsername	Returns the accounts results in a search for a given username
generateHeaders	Adds credentials to a query
getMedias	Retrieves the n most recent posts for a given username
getMediaById	Retrieves information about a post with a given ID
getMediaByUrl	Retrieves information about a post with a given URL
getMediaByCode	Retrieves information about a post with a given shortcode
getPaginateMedias	Returns the first page of recent posts for a given username
getMediaCommentsById	Retrieves comments on a post with a given ID
getMediaCommentsByCode	Retrieves comments on a post with a given shortcode
getMediaLikesByCode	Retrieves likes on a post with a given shortcode
getAccountById	Retrieves information on an account with a given ID
getAccount	Retrieves information on an account with a given username
getMediasByTag	Retrieves n most recent posts with a given hashtag
getPaginateMediasByTag	Returns the first page of recent posts with a given hashtag
getCurrentTopMediasByTagName	Retrieves the top nine posts with a given hashtag
getCurrentTopMediasByLocationId	Retrieves the top nine posts for a given location ID
getMediasByLocationId	Retrieves the n most recent posts for a given location ID
getLocationById	Retrieves information for a given location ID
getFollowers	Retrieves the accounts following a given account ID
getFollowing	Retrieves the accounts a given account ID is following
getStories	Retrieves the story information for a given account
setProxy	Establishes a proxy for queries to be passed through

`instaR`, `Instagram-php-scraper` accesses Instagram's open API. This allows users to retrieve most public information without needing authorization from Instagram or even an account. However, it does allow for users to log-in to access information limited to authenticated accounts. The functionality of this program is also much more extensive than others. Note, however, that while it has functions that interact with specific locations, it cannot find posts in a given region. The two main factors that make

`Instagram-php-scraper` an imperfect candidate is the language it is coded in, PHP, and that it is controlled by a source outside the DoD. It does, however, serve as an excellent blueprint for other programs interacting with Instagram's open API.

2.6 Reproducible Analysis

The last subject that this thesis aims to address is the reproducibility of analysis. To the ends of validation and verification, is has been standard for independent bodies to try and replicate the results of a study (R. D. Peng, 2011). However, with more research being done with the aid of data and independently collected information, all sciences are suffering a crisis of reproducibility (R. Peng, 2015). Pulling information from a website can result in drastically different outputs if the procedures for searches and search criteria are not specified. Even then, the information might have changed since most websites allow edits and deletions, as well as new data always being entered. The concept of ensuring that the data and procedures for a study are made available to others is referred to as "reproducible research" (R. D. Peng, 2009). For the context of intelligence agencies, this is valuable as an extra step of validation. One of the largest benefits of using open source software is that it is highly reliable and reproducible (Ven & Verelst, 2006). When future users attempt to reproduce the results of previous work, open source programming languages like R allow them to ensure the same work is being done. In a practical sense, this means that a user knows what to expect when they run a function. Unlike in closed source software, a user could even look at the source code to verify that the same operation is being carried out. In scholarly and scientifically rigorous applications, developers should design software to facilitate reproducibility.

III. Methodology

3.1 Overview

The goals of this research are achieved by creating an R package called *instaExtract* hosted on the GitHub page for the Data Science Lab headquartered in the Air Force Institute of Technology. The package uses publicly available links from Instagram to retrieve information on users, posts, and locations. Using R as the backbone for this package allows analysts to easily handle the data returned as well as create reproducible reports. Being hosted on the Data Science Lab's GitHub page allows for the package to be both readily accessible and routinely updated. As a demonstration of the future of such a package this thesis includes an example function, `createLocationMapping()`, to demonstrate what a future user might contribute to the project.

3.2 A JSON-Based Instagram Scrapper in R– `instaExtract`

The limitations of other accessible Instagram scraping software have already been discussed. In addition to being outside the military community's reach of control, these other programs suffer from challenges that make them unsuitable for use by intelligence gatherers. Namely, they are limited by interacting with the restrictive and exclusive Instagram API or they are difficult to operate and distribute.

This research overcomes these challenges with the creation of an R package that is based on interactions with publicly available Instagram links that return JSON files. The primary advantage of acquiring data this way is that most information on the site can be obtained without the need to log-in, nor interact at all with the closed API. This would allow a user behind a proxy to access almost all information on Instagram without

needing to provide any of their own information. Furthermore, the use of JSON links is far less demanding on a user's bandwidth than loading every page in a traditional browser. A JSON link will not load images or videos which greatly reduces the time demand of queries. Instead, all the information for a page is reduced to the data defining it. For instance, an image would be represented by a unique identification number and Uniform Resource Locator (URL). If, in the future, a user wanted to download an image, they could use this information to download it.

3.2.1 Helper Functions

There is a collection of generic links that the package uses to create these JSON links. To access the information a user wants, the package will take a user input and substitute it into the generic link for each category. For more complex interactions, the package will use information contained in the JSON file to create a new link, simulating pagination through all the possible results, or until the desired number of results is returned. These links and their manipulators are located in the `Endpoints.R` file. In the example below, an example user query is used to create a working link to Instagram's open API.

```
User_Media_Json_Link <-
"https://www.instagram.com/{username}/?__a=1&max_id={maxID}"


getUserMediaJsonLink <- function(user, maxID){
  link <- User_Media_Json_Link
  link <- gsub("{username}", user, link, fixed = TRUE)
  link <- gsub("{maxID}", maxID, link, fixed = TRUE)
  return(link)
}
```

To work with Instagram's identification numbers, the package utilizes transformer functions in `Transformers.R`. Instagram uses both IDs and shortcodes to identify their content. IDs are most commonly returned in the JSON file, but shortcodes are the identifiers used when creating the links. To go from shortcode to ID, for each character in the code, the program multiplies a placeholder ID by 64 and adds that number to the index of the given character in an alphanumeric alphabet. To go from ID to shortcode, process is reversed. However, since Instagram's IDs are larger than the integer size that R works with, the package utilizes the `bit64` package to accurately represent the integer when doing math. The functions used to perform these operations are shown below.

```
getIDFromCode <- function(code){
  alphabet <-
  "ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-_"
  ID <- 0
  for(i in 1:nchar(code)){
    c <- substr(code,i,i)
    id <- id * 64+ regexpr(c, alphabet)[1]
  }
  return(ID)
}


getCodeFromID <- function(ID){
  alphabet <-
  "ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-_"
  code <- ""
  ID <- bit64::as.integer64.character(ID)
  while(ID >0){
    remainder <- ID %% bit64::as.integer64.double(64)
    ID <- (ID - remainder) %/% 64
    code <- paste(substr(alphabet,remainder+1,remainder+1),code,sep="")
  }
  return(code)
}
```

3.2.2 Search Functions

With endpoints and transformers established, the package can utilize a large assortment of functions to pull specific information or search for a user provided query. The smaller subset of functions are the search functions, which return tags or account names similar to the query given by the user. Table 4 shows these two functions and their purposes.

Table 4 - Search Functions Provided by the `instaExtract` Package

Function	Purpose
`searchAccountsByUsername`	Will return a data frame of accounts and their information whose username contains the users query
`searchTagsByTag`	Will return a data frame of tags in use that contain the users query

3.2.3 Get Functions

The majority of `instaExtract`'s functionality comes in the form of its Get functions, which retrieve information based on what the user supplies. Table 5 lists these functions as well as their purposes. All these functions operate without a user needing to log in. When using a function, the user will typically provide two arguments, the query they are interested in and the number of results they wish to return. In cases where the number is not provided, the package will only return the first page of information. The package will then return a data frame containing information relevant to the given element in question. These results are detailed in Table 6.

Table 5 - Get Functions for the `instaExtract` Package

Function	Purpose
`getCommentsByMediaCode`	Return the first n comments and comment information for a media with a given shortcode
`getCommentsByMediaID`	Calls getCommentsByMediaCode after transforming a media ID into a shortcode
`getCurrentTopMediaByLocationID`	Returns the top nine media for a location with the given location ID
`getCurrentTopMediaByTag`	Returns the top nine media for a given hashtag
`getLikesByMediaCode`	Returns information on the first n likes on a media post with the given shortcode
`getLikesByMediaID`	Returns information on the first n likes on a media post with the given ID
`getLocationByID`	Returns information on a location for a given location ID
`getMediaByCode`	Returns information on a media post for a given media shortcode
`getMediaByID`	Calls `getMediaByCode` after transforming ID to a shortcode
`getMediaByLocationID`	Returns information on the n most recent media post for a given location ID
`getMediaByTag`	Returns information on the n most recent media post for a given hashtag
`getMediaByURL`	Returns information on a media post with a given URL
`getMediaByUsername`	Returns the n most recent media posts from an account with the given username

Table 6 - Format of the Results for the Get Functions in the `instaExtract` Package

Function	Results
`getCommentsByMediaCode`	n x 6 data frame: id, text, created_at, owner.id, owner.profile_pic_URL, owner.username
`getCommentsByMediaID`	n x 6 data frame: id, text, created_at, owner.id, owner.profile_pic_URL, owner.username
`getCurrentTopMediaBy LocationID`	9 x 14 data frame: id, thumbnail_src, tubmnail_resources, is_video, code, date, display_src, video_views, caption, dimensions.height, dimensions.width, owner.id, comments.count, likes.count
`getCurrentTopMediaByTag`	9 x 14 data frame: id, shortcode, taken_at_timestamp, display_URL, tumbnail_src, tumbnail_resources, is_video, video_view_count, edge_media_to_caption.edges, edge_media_to_comment.count, dimensions.height, dimensions.width, edge_liked_by.count, owner.id
`getLikesByMediaCode`	n x 7 data frame: id, username, full_name, profile_pic_URL, is_verified_followed_by_viewer, requested_by_viewer
`getLikesByMediaID`	n x 7 data frame: id, username, full_name, profile_pic_URL, is_verified_followed_by_viewer, requested_by_viewer
`getLocationByID`	1 x 6 data frame: id, name, has_public_page, lat, lng, slug
`getMediaByCode`	1 x 42 data frame: meta information about a post not including its comments and likes
`getMediaByID`	1 x 42 data frame: meta information about a post not including its comments and likes
`getMediaByLocationID`	n x 15 data frame: comments_disabled, id, thumbnail_src, thumbnail_resources, is_video, code, date, display_src, video_views, caption, dimensions.height, dimensions.width, owner.id, comments.count, likes.count
`getMediaByTag`	n x 16 data frame: comments_disabled, id, thumbnail_src, thumnail_resources, is_vdeio, code, date, display_src, caption, dimensions.height, dimensions.width, owner.id, comments.count, likes.count
`getMediaByURL`	1 x 42 data frame: meta information about a post not including its comments and likes
`getMediaByUsername`	n x 17 data frame: _typename, id, comments_disabled, getting_info, media_preview, thumnail_src, thumbnail_resources, is_video, code, date, display_src, caption, dimensions.height, dimensions.width, owner.id, comments.count, likes.count

In order to provide a better understanding of how the package operates, this paper provides a walk-through of the `getMediaByTag` function, which is representative of

how most of the other Get functions operate.

```
getMediaByTag <- function(tag, n = 20, maxID = ""){
```

The function call has three arguments: 1) `tag` – the query a user is looking to gather data on; in this case, a hashtag. 2) `n` – the number of responses that will be returned. Each function defaults to the number of results that are returned in the first JSON file. 3) `maxID` – an identifier that allows the function to start its search from a different position. The user will usually only use the default of an empty character, but the function will use it to paginate through JSON results.

```
 i <- 0
 moreAvailable <- TRUE
 data <- data.frame()
```

Next, the function initializes variables used in its operation. The integer `i` is used to keep track of the number of results returned. The boolean `moreAvailable` will track if more results from Instagram are available. Initializing data as a data frame allows inserting new rows as results come in.

```
#will run while more data exists and it has not reached n results
while(moreAvailable && i < n){
```

The main operations of the function, creating a link and collecting the results, will happen in this while loop. It will loop until there are no more results available or until the number of results is greater than `n`.

```
   #create the url from Json Link
    url <- getTagMediaJsonLink(tag,maxID)


Tag_Media_Json_Link <-
"https://www.instagram.com/explore/tags/{tag}/?__a=1&max_id={maxID}"


getTagMediaJsonLink <- function(tag, maxID){
  link <- Tag_Media_Json_Link
  link <- gsub("{tag}", tag, link, fixed = TRUE)
  link <- gsub("{maxID}", maxID, link, fixed = TRUE)
  return(link)
}
```

Calling the helper function `getTagMediaJSONLink` creates the JSON link URL. When this function is called, it passes the desired hashtag and the current `maxID`. The helper function will then alter a constant string that was defined before compilation by inserting the given hashtag and `maxID`. The character string this function returns is the URL that will be used to retrieve the desired JSON file.

```
      #the unflattened response
      response <- jsonlite::fromJSON(url)


      #will return as list if there is only one result


      if(!is.data.frame(response$graphql$hashtag$edge_hashtag_to_media$edges$
      node)){
            return(response$graphql$hashtag$edge_hashtag_to_media$edges$node)
          }


          else{
            #flattening the data down to the nodes, into a dataframe
            media <-
      jsonlite::flatten(response$graphql$hashtag$edge_hashtag_to_media$edges$
      node)
```

Next, the URL is used to retrieve the JSON file from Instagram. The imported function `jsonlite::fromJSON` translates the JSON file to a R usable data frame. If there is only one result in this JSON file, it will return this list without doing the rest of the function which assumes that it has a data frame. In the normal case, when a data frame is returned, `jsonlite::flatten` will flatten the data frame to be a regular two dimensional data frame, meaning that each column of the data frame is atomic, rather than a list with multiple values. In most cases, the desired information is located inside containers such as "graphql" and coupled with superfluous information. To target the useful information, the package must reference the correct container in the data frame returned from the JSON link. In each case, this path is unique to a function and can change between Instagram updates. This points towards a need to have this package quickly modifiable and subsequently distributable, which is addressed in Section 3.3.

```
    #iterating over the rows of the media
    for(row in 1:nrow(media)){


      #will exit loop and return data if reaching the limit
      if(i == n){
        return(data)
      }


      #will add a new row of media to data
      data <- plyr::rbind.fill(data,media[row,])


      #incrementing the counting index
      i <- i + 1
    }
```

This segment of code will iterate through the rows of the results returned from Instagram. Each time, it will add that row to the data frame intended to be returned. The `rbind.fill` function from the `plyr` package to ensures that information from video

and picture posts can be added to the same data frame. The function keeps track of how many rows it has entered and will return the results if it reaches the desired number.

```
      #Where to start the next query to the instagram link
      maxID <- response$graphql$hashtag$page_info$end_cursor
      #makes sure more exists
      moreAvailable <-
response$graphql$hashtag$page_info$has_next_page
    }
  }
```

The last step in the while loop is storing the identifier that allows the function to retrieve the next set of information. Again, emphasis is given to the nature of the JSON files' structures returned. The location of these identifiers inside the JSON is subject to change, and without a method for updating, the function might become obsolete.

```
#convert the json data to R dataframe
  return(data)
}
```

Now that data is filled with n results, it is returned to the user. Each Get function returns a variety of data frames with different elements, as outlined above in Table 6. The remaining Get functions operate very similarly to getMediaByTag, however each has a unique URL and data structure to its JSON file.

The Get functions in this package represent most of the functionality available in other Instagram scrapping programs. Not currently possible is retrieving following and follower information from an account, or getting a username from an owner ID. This gap in functionality also means that the package operates without needing log-in credentials or interaction with the Instagram API. Utilizing the JSON-based open API, the package

relies on constants that must be defined by the developer. Therefore, the maintainer of the package is incentivized to house it in a manner that is adaptable to updates to Instagram. This concern is solved by hosting the package on the AFIT Data Science Lab's GitHub page, a process discussed in the next section.

3.3 Adaptable, Reproducible, Distributable – Hosting on the AFIT Data Science Lab's GitHub Page

To ensure the `instaExtract` package is adaptable, reproducible, and distributable, it is hosted on it on the GitHub page of the AFIT Data Science Lab. From their website, "GitHub is a code hosting platform for version control and collaboration" ("Hello World · GitHub Guides," n.d.). That succinct description captures why GitHub can be an indispensable tool for this package and for many military analysis operations. GitHub allows the package to be accessible to all systems that have access to the internet, to store previous versions of the package to reproduce reports done in the past, and to facilitate discussion and evolution between users and developers of the package.

3.3.1 Benefits of GitHub for Military Analysts

An immediate consequence of hosting the package in an online repository is that the most current edition of the package is available to any user with access to the internet. The tools needed to conduct Instagram data scraping are not locked away on a machine. Instead, the user can have a consistent and reliable way to locate the tools they are familiar with and confident in. A further benefit of a centralized hub like the AFIT Data Science Lab is that consolidated resources will lead to more capable and competent tools. When the standard practice for analysts becomes working from a common set of applications, those applications benefit from increased attention and use. Additionally,

users who begin working within established systems will have increased code and data compatibility. The benefits of consolidation and cooperation are multiplicative.

GitHub furthers this research's goal of enabling reproducible analysis. R as a language already facilitates reproducibility with workable data types and dynamic reports in R markdown. GitHub goes a step further and ensures that the exact version of the package used to construct the report is available to an analyst who is validating it. GitHub stores the changes made to a package, meaning as long as the version of the package used is recorded in the report, it can be downloaded on the GitHub page. This prevents future changes to a function from impacting the validation process. An older version of the package should operate just as it did when the analysis was first conducted.

Perhaps the strongest benefit of using the GitHub platform is the level of collaboration and adaptability. It is a common occurrence in large organizations, the military included, for there to be a large disconnect between the developer of a tool and the end users of that tool. The end user is often left with needless capabilities or lack those that are essential. GitHub facilities a connection between users and developers, and in some cases, blurs the distinction between the two. The development process for a conventional program hosted on a local machine is filled with inefficiencies and difficulties. First, the developer is made aware of a need. This alone might have taken months to perforate a command structure and reach their desk. Next, they need to collect information on the end user's needs or operate off of the limited information they might have been given with the assignment. After working on the project for a time, they will seek feedback with the end user. This will lead to revision cycles that take time, money and are bound to miss complications that will arise when the program ships. The resulting

program will then need a method of distribution, which can range from an online download, to an executable passed through a file sharer. Whatever the method, the end user is left with a program that might or might not meet all their needs. The first few weeks of operations will be monitored by the developer, who can make necessary changes and encourage users to download the most recent version. But after a time, the gap between a program's users and the developer widen. New issues arise, whether from updated needs of the user or from unforeseen use cases. Complaints may be raised, but resolutions suffer from the same slowdowns that the initial development suffered from. If users result to fixing these issues themselves, different versions of the program can begin to diverge, causing new issues with incompatibility and distribution. The end result of the conventional set up is that developers are isolated from the users they are creating for and the users are stuck with programs that are stagnant and restrictive.

Comparing that process with that of development through GitHub illustrates the benefits of a shared hosting platform. With centralized hubs of development, like the AFIT Data Science Lab, requests for new tools can be addressed with expedience. Depending on the scope and the urgency of the problem, the development might be conducted by students, civilians (government or non-government), or dedicated military analysts. In all cases, the development process can be much more organic than the conventional process. Errors and inadequacies are to be expected to some degree in all forms of computer programs, but GitHub facilitates communication and issue resolution to minimize the distribution flaws cause the user. In the conventional example, an issue that arises after the development process of an application can be hard to remedy. With GitHub, the end user can instantly notify the developer or the staff assigned as

administrators to the application. The error might be as simple as misuse of the application, in which case, the administrator can reply to the issue and guide the user through operation. If the issue turns out to be something more complex, the administrator can use the issue as immediate feedback about the workings of the application. Since the application is hosted in one predictable location, any updates the administrator makes to fix the problem will automatically be reflected in the current of the program. Users can always ensure that their programs are up to date and they are operating with the same tools that everyone is using.

3.3.2 An Example of GitHub's Potential Adaptability – Location Mapping in the `instaExtract` Package

GitHub also offers users the ability to directly submit their own contributions, blurring the line between developer and user and greatly increasing the adaptability of a program. If a user finds that they have need for a new function, its within their power to create it. GitHub gives user access to the source code, allowing them to add to and modify their local version of the package. More than that, GitHub provides the means for that user to add their own custom function to the main package. They may request a change to the main program, and provided the administrators approve the change, that custom function can now be used any other user with access to the package. This integration of users into the development process results in ever expanding functionality and increased resiliency.

To illustrate the potential of user created functionality, the `instaExtract` package includes a set of functions that broaden the capability of the package and enable users to conduct more thoughtful analysis. This set of functions is detailed in Table 7. A

current weakness in the JSON dependence of the package compared with the Instagram API is that there is no readily available way to search for posts around a given latitude and longitude. The solution presented by this research to this problem relies on extracting the latitude and longitude from a location's page. These values are then used to create a location mapping, a data frame with locations and their information.

Table 7 - Functions Enabling Location Mapping Capabilities

Function	Purpose
`createLocationMapping`	Creates a data frame with all locations in a search region
`haversineDistance`	Calculates the distance in miles between two latitude and longitude points
`getCurrentTopMediaByLocationMapping`	Returns the top n results for each location in a location mapping
`getMediaByLocationMapping`	Returns the n most recent results for each location in a location mapping
`getLocationsInRange`	Returns a filtered location mapping with locations within a certain radius

The core function developed to achieve these functions is called `createLocationMapping`.

```
createLocationMapping <- function(country ="", city = "", lat_long = FALSE){
```

This function creates the data frame in the format referred to as a "location mapping" in this paper. It has three arguments: 1) `country` – the name/s of a country to search inside. If no name is given, the mapping will be for all countries in Instagram's explore page. 2) `city` – the name/s of a city to search inside. If no name is given, the mapping will be the top 1000 cities for all countries in the scope of the function. 3)

`lat_long` – if true, the function will gather the latitude and longitude values for each location. It is defaulted to false since this operation requires a web query for each location and can take significant amounts of time.

The key component to the operation of the `createLocationMapping` function is the explore/locations page on Instagram. The top layer for this subdomain lists the names of all countries in Instagram's databases. The function will grab the names of all countries in its scope, as well as their slug's and ID. Using this information, it can move down a layer to the country's subpage, where it lists the top 1000 cities for that country. Likewise, the function will grab information from this page to travel to each city within the argument's scope. The last level of explore/locations is a list of the top 1000 locations in all the selected cities. These locations are what are added to the location mapping, along with their slug and ID and the slug and ID of the city and country they belong to. If the user selected to gather coordinates from the location, the function will also visit each location page to grab the latitude and longitude. This can take a very long time depending on the speed of the user's internet connection and the number of locations in the location mapping. However, since the information is retained in a data frame, a user can utilize location mappings made in prior, avoiding this time constraint. Furthermore, it would be possible for future contributors to build package that facilitate this process by hosting current versions of location mappings. This would mean the analyst only has to download the relatively small file instead of performing thousands of time consuming queries.

With the location mapping created, it offers the user new manners for conducting smart analysis on Instagram's potential data. Two of the functions extend the package's

ability to get media from a location to getting media from a location mapping. The `haversineDistance` function enables the user to filter a location mapping to only locations with a mileage radius of place of interest. To demonstrate what these functions add to the package, this paper walks through an example scenario. Suppose that a user is interested in what posts are being made in vicinity of the White House. The first step would be to create the location mapping of Washington D.C.

```
washington_dc <- createLocationMapping("United States", "Washington", TRUE)
```

This function call creates the a mapping for the city "Washington" in the country "United States" with the coordinates information. The variable `washington_dc` now contains information for the top 1000 locations on Instagram in Washington DC.

```
lat <- 38.8977
long <- -77.0365
range <- .5 #miles
```

Next, the user sets the region they are interested in. The coordinates for the White House are 38.8977° N, 77.0365° W. In this case, they are interested in locations within half a mile from the point.

```
locations_near_white_house <- getLocationsInRange(washington_dc, range, lat,
long)


getLocationsInRange <- function(mapping, r, lat, long, ...){
  if(!is.numeric(r) || !is.numeric(lat) || !is.numeric(long)){
    stop("r, lat, and long, must be numeric")
  }
  mapping <- filter(mapping, haversineDistance(latitude, longitude, lat,
long) <= r)
  return(mapping)
}
```

Using the simple function, `getLocationInRange`, the user can narrow down the location mapping to a location mapping of only those locations within the range of the white house set before. The `haversineDistance` function will return the distance of each point from the White House, and only locations within half a mile from the White House will be included in the location mapping named `locations_near_white_house`.

```
near_house_media <- getMediaByLocationMapping(locations_near_white_house)
```

The last step is to use location mapping version of the get media function to retrieve information on the recent media from the locations within our range. This procedure results in a data frame with information on the most recent posts made at locations near the White House. A user would be free to conduct whatever analysis they wish with the data provided, but possible areas of interest might be mapping commonly used hashtags for an area, comparing data from multiple days to observe fluctuations in time, identifying prolific accounts, etc.

The location mapping functions created represent the future potential of the `instaExtract` package or any package hosted on a collaborative platform like GitHub. As users discover new ways to conduct analysis or require new data sources, the functions they spur or create themselves can be added to the package. This in turn, facilitates and inspires other analysts in conducting sophisticated and modern analysis of their own.

3.4 Summary

This research culminates in a R package called `instaExtract` hosted on the AFIT Data Science Lab's GitHub page. R is an open source program that is easily accessible and built around reproducible practices. The `instaExtract` package provides users with a way to scrape current Instagram data without authentication or interaction with the Instagram API. By hosting the package on the AFIT Data Science Lab's GitHub page, the package has an increased potential to stay relevant and useful to real world analysts. Also included is an original set of functions curtained around mapping locations to illustrate how future collaboration on this project might continue to extend its usefulness.

IV. Exemplar Analysis

4.1 Overview

To better illustrate the capabilities and value of the `instaExtract` package, this thesis includes an exemplar analysis that will emulate how an end user might use the package. The user is assumed to have a working understanding of R and Instagram. The first scenario focuses of an analyst investigating threats near a location, in this example the White House. Section 4.2 discusses the gathering of information about the area. Next, the analysis will conduct basic filtering and visualization techniques as well as other `instaExtract` functions to investigate further.

4.2 Gathering Information on the Region

The investigation of the threat begins by collecting information about the locations surrounding the White House. The user creates a location mapping of the Washington D.C. region. The first 30 entries of the data frame of 1000 locations generated by this function are found in Appendix B.

```
washington_dc <- createLocationMapping("United States", "Washington", TRUE)
```

Next, the location mapping is narrowed down to find only the locations within a mile of the White House. As in the demonstration in the methodology section, the user sets the coordinates of the White House and use the function `getLocationsinRange` to filter the locations. The first 30 entries of the data frame of 524 locations are shown in Appendix C.

```
lat <- 38.8977

long <- -77.0365

range <- 1 #miles

locations_near_white_house <- getLocationsInRange(washington_dc, range, lat,
long)
```

Again, the next step is to collect the media from the area, but to be a little more thorough, this user will get the last 50 posts from each area. The first 60 of 26,200 results are partially shown in Appendix D. The complete values of the URLs and captions are obscured to allow the table to be represented on one page. This operation took about 30 minutes, meaning that `instaExtract` is able to pull the information about a post around 15 times a second.

```
media_near_white_house <- getMediaByLocationMapping(
locations_near_white_house, 50)
```

4.3 Hashtag Investigation

Now possessing media data, the user can begin to investigate it. There are many different avenues to take at this point, where the following courses represent only some of the possible paths. First, the user chooses to investigate the hashtags being used. The hashtags contained in the captions are isolated and sorted to find the most commonly used hashtags. Using the `tidytext` and `stringr` packages, the user first breaks down each caption to separate elements for each word. Then, the words are filtered down to only those words that start with '#'; our hashtags. Sorting these, the user finds the most common hashtags recently used in the search region. The results indicate that there are 36971 hashtags used in total, with the top 15 displayed below. Filtering that result further, only 59 hashtags have been used more than 100 times in search radius. Further,

as Figure 5 and Figure 6 show, even the top 150 hashtags are greatly skewed to the top, with the top two, “#washingtondc” and “#dc” being used upwards of 2000 times each, with next highest, “#washington” only being used 513 times.

```
library(stringr)
library(tidytext)


hashtags <- media_near_house3 %>%
  unnest_tokens(word, caption,token = "regex") %>%
  filter(substr(word,1,1) == '#') %>%
  count(word,sort = TRUE)

# A tibble: 15 x 2
                  word     n
                 <chr> <int>
 1       #washingtondc  2612
 2                 #dc  2171
 3         #washington   513
 4             #travel   489
 5                #usa   430
 6               #igdc   423
 7               #love   405
 8          #instagood   371
 9        #acreativedc   369
10                #tbt   350
11        #photography   338
12             #foodie   325
13           #foodporn   313
14 #washingtonmonument   296
15      #valentinesday   295
```

Figure 5. Word Cloud for Top 200 Hashtags in the DC Area

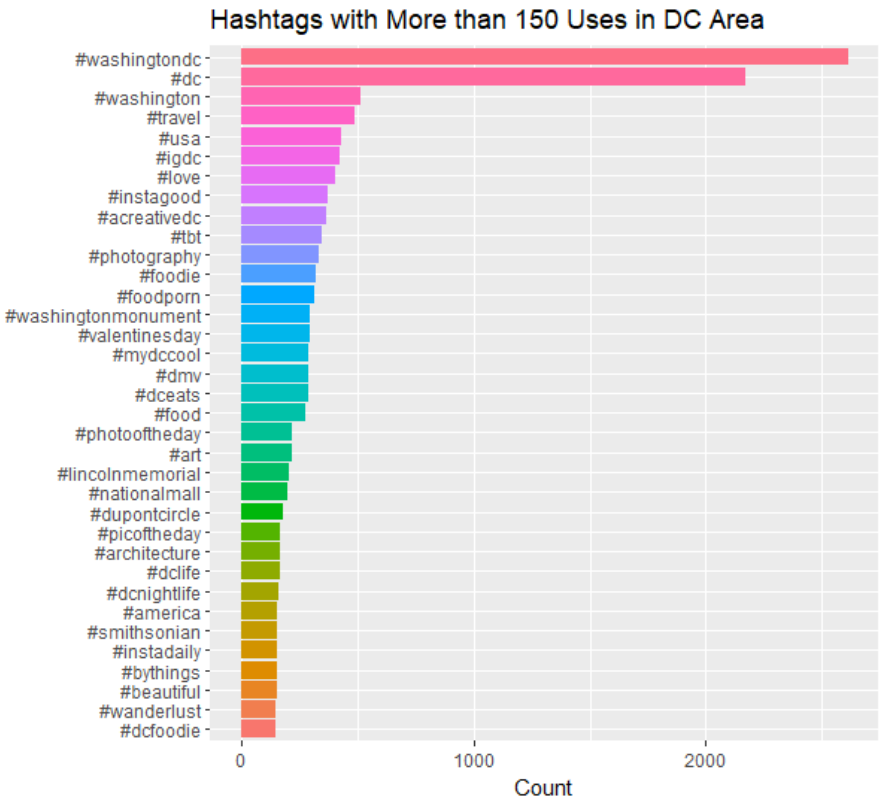

Figure 6. Count of 150 Most Used Hashtags in the DC Area

For the next data utilization scenario, the user has a list of key words that they are monitoring. It is possible to compare the list of keywords against the list of hashtags to look for suspicious hashtags. The code to perform this comparison and the top 6 results of such a search are shown below. It reveals 148 potentially threating hashtags. The top result, “#justgoshoot”, has been used 45 times in the scope of the search.

```
keywords <- c('shoot', 'gun', 'kill', 'bomb')


threats <-hashtags %>%
  filter(str_detect(word, paste(keywords, collapse = "|")))

# A tibble: 6 x 2
                   word     n
                  <chr> <int>
1          #justgoshoot    45
2           #photoshoot    34
3           #shoot2kill    10
4          #citykillerz     7
5 #shootingwithshooters     7
6          #tonekillers     7
```

This hashtag turns out to be a hashtag used by photographers, but if it really was a threat, `instaExtract`can continue provide more information. To find more data on the usage of this hashtag, the user retrieves the last 10,000 posts that used this hashtag. Further, to know who used this hashtag most in the most recent 10,000 uses, the user counts the list again, and finds that the account with ID 6961358493 used the hashtag 50 times. Outside tools could provide for even more complex analysis. The large amount of written word in the captions makes a good source for sentiment analysis. This data contains numbers of likes and comments, so a user could also sort for popular posts that fit a given criteria.

```
just_go_shoot <- getMediaByTag("justgoshoot", 10000)


just_go_shoot %>%
  count(owner.id, sort =TRUE)


owner.id     n
        <chr> <int>
 1 6961358493    50
 2 2014353892    45
 3  495504345    37
 4 7129727912    30
 5   29197162    29
 6 4843411210    29
 7  374142894    28
 8  298064835    24
 9 1524025234    20
10 4976873299    18
```

4.4 User Investigation

A user can pull information from an account to find more detailed information about that account's history. The first step would be to download media from the chosen account's page. In this case, the account under investigation has the username "usairforce." Retrieving their last 2000 posts returns 1390 values, the first 60 of which are shown in Appendix E, since this account has only posted 1390 times.

```
usairforce <- getMediaByUsername("usairforce", 2000)
```

The data collected can be used to get a better idea user's is attributes. The package user begins by finding the post with the most likes, their most popular post. Running any code is not needed to do this, as R's data viewer allows sorting a column in descending value. In this case, the account's top post, ID 1698060742908836418, has 41,541 likes, and the photo is shown in Figure 7. The number of likes plotted against the postdate results in

Figure 8, showing the trend of like counts over time. This figure which shows a gradual increase in the average likes over time.

Figure 7. Most liked Instagram Post by usairforce (usairforce, 2018)

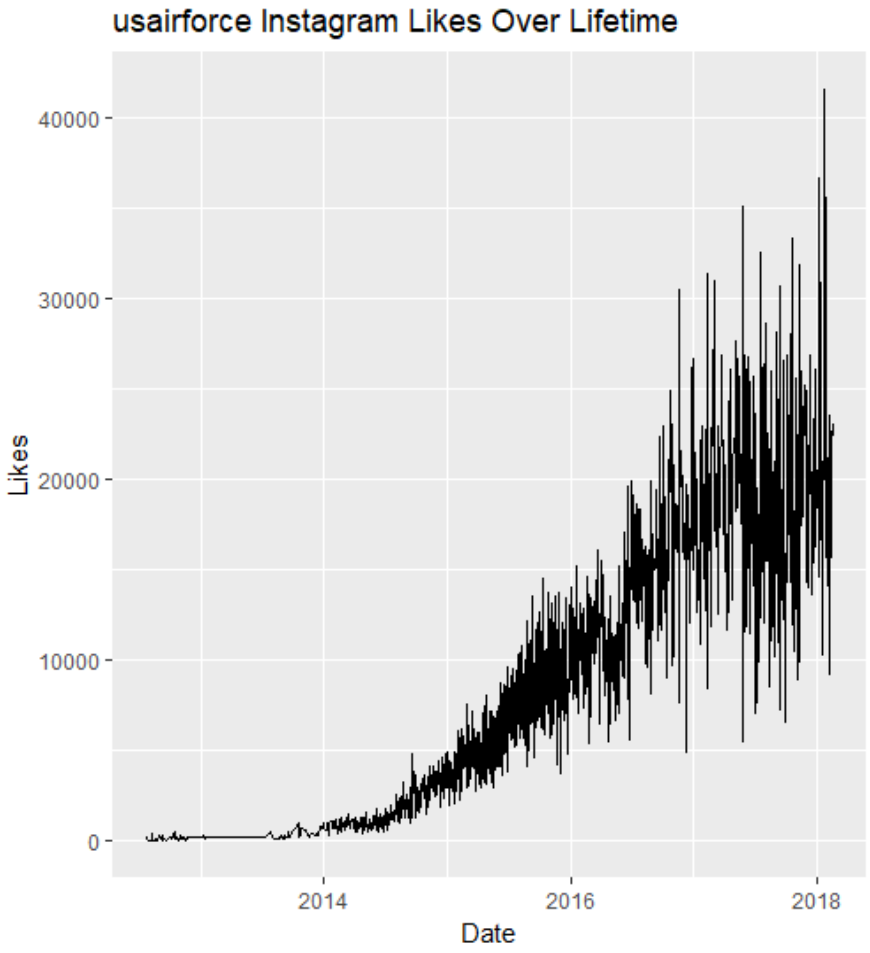

Figure 8. Post Likes Over Time for usairforce Account

The package also allows a closer look at a post. A media ID is used to retrieve the comments from the top post on this account in the `getCommentsByMediaID` function, the first 60 of 269, which are shown in Appendix F. Using this information, the most common words used in the captions are found in a process similar to that used above.

```
comments <- getCommentsByMediaID('1698060742908836418', n = 300)


commentWords <- comments %>%
unnest_tokens(word, text,token = "regex") %>%
filter(!word %in% stop_words$word) %>%
count(word,sort = TRUE)%>%
mutate(word = reorder(word, n))

# A tibble: 6 x 2
      word     n
    <fctr> <int>
1     love    14
2      air    13
3 rdgjklcb    11
4    force    10
5     girl     7
6    pilot     7
```

Comments can also provide insight to connections between users. Using the comments from posts, the user can build a network of users who interact with the usairforce account. To begin, 300 comments from the last 50 posts are collected. From these, the users who have left the most comments is calculated. Although simple, the network built from this data, Figure 9, hints at how networks can be developed from the social media data obtained this way.

```
all_comments <- data.frame()

for(row in 1:50){
  comment_holder <- getCommentsByMediaID(usairforce[row,'id'],300)

  all_comments <- plyr::rbind.fill(all_comments, comment_holder)
}


commenting_users <- all_comments %>% count(owner.username, sort=TRUE)


# A tibble: 6 x 2
               owner.username     n
                        <chr> <int>
1 my_babel_physics_project_ai    30
2             admininnotinuse    28
3                zafarwestern    26
4             gomezbaquerosol    21
5        viktoria__lg.troxell    21
6               lydiavassallo    20
```

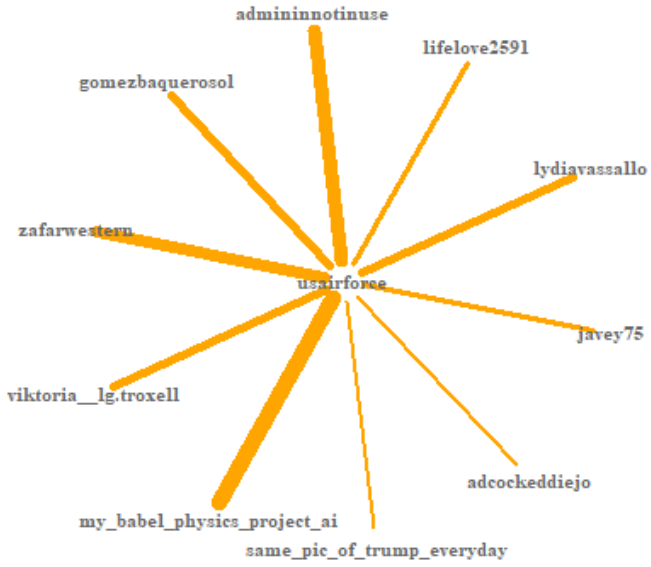

Figure 9. Network of Commenting Users on usairforce Posts Where Edge is Proportional to Number of Comments

4.5 Geo Mapping

Part of the analytic cycle, shown in Figure 4, is visualizing the data. This allows the analyst to grasp what type of data they are working with and understand the significance of their analysis. R has many great packages that can help an analyst accomplish this feat. One of these packages is `leaflet`, a JavaScript library used to make interactive maps (Cheng, 2017). This section revisits the data used in section 3.3.2, the location mapping of Washington DC. Using `leaflet`, an analyst can overlay the locations in the location mappings with a real map of the area. The result, seen in Figure 10, is cluttered beyond much helpfulness.

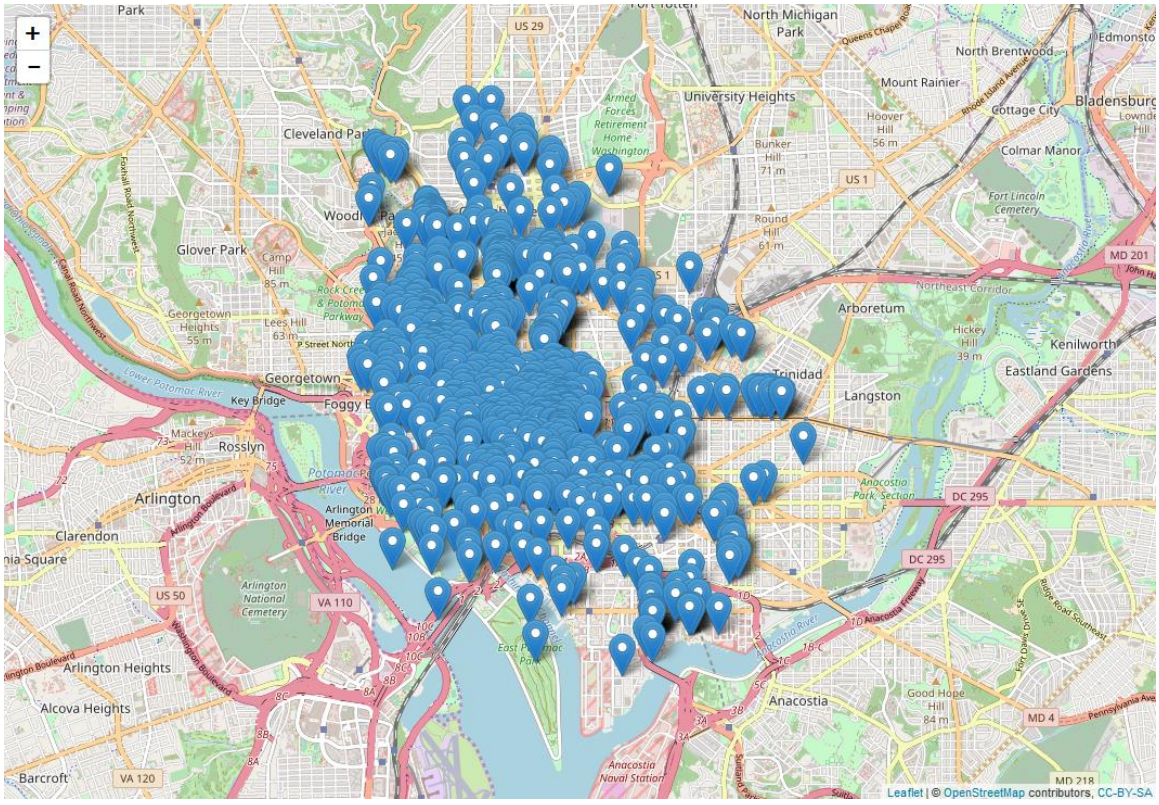

Figure 10. All Washington DC Locations from Instagram Overlaid with Actual Location

Luckily, `leaflet` comes with a clustering option that drastically increase the readability of the map. In Figure 11, clustering has been turned on. The analyst can now see where the locations are focused, without overwhelming the senses. Hovering over a cluster shows the region it represents. For closer detail, the analyst can click on a cluster to zoom in and expand its contents. Once a node is clicked, leaflet will display the name of the location. Figure 12 is a fully zoomed in example. All of the data displayed in these maps is nothing more than the information in a location mapping, but displaying that information in a real world setting adds context and meaning to the information that allows analysts to draw connections and make inferences that they would be hard pressed to do without that context.

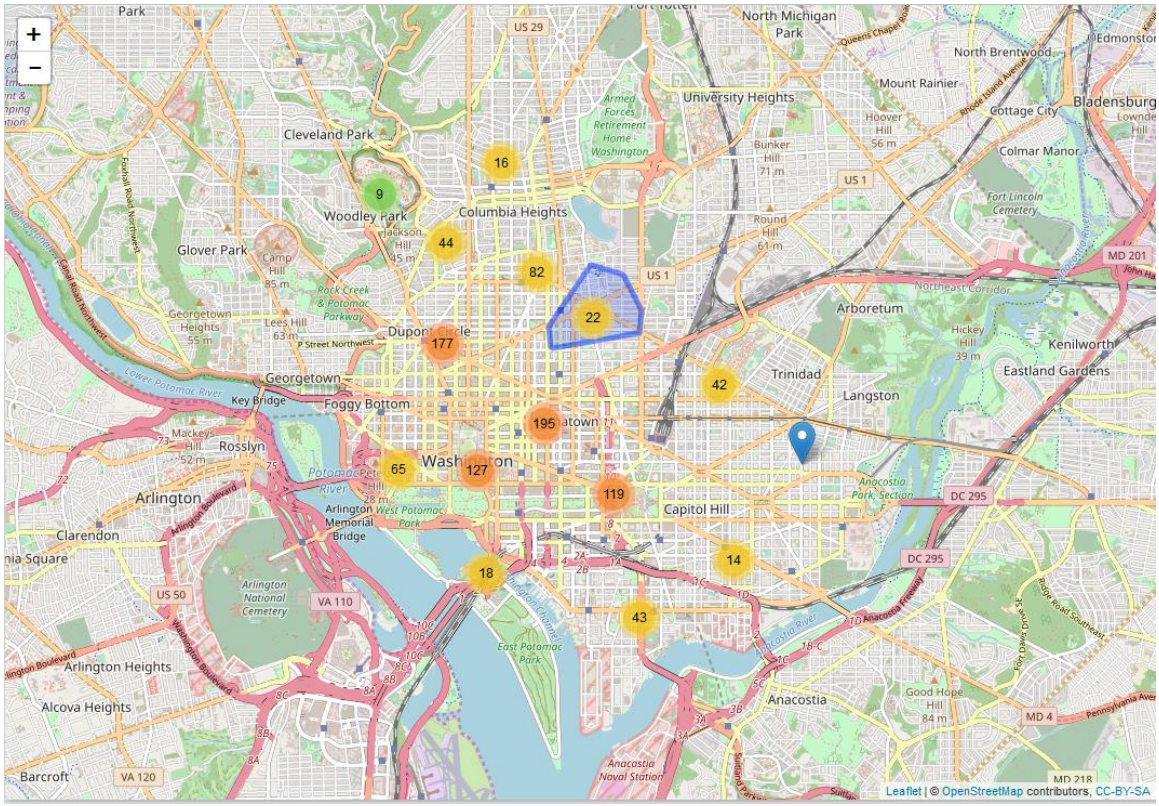

Figure 11. Clustered Locations for Washington DC

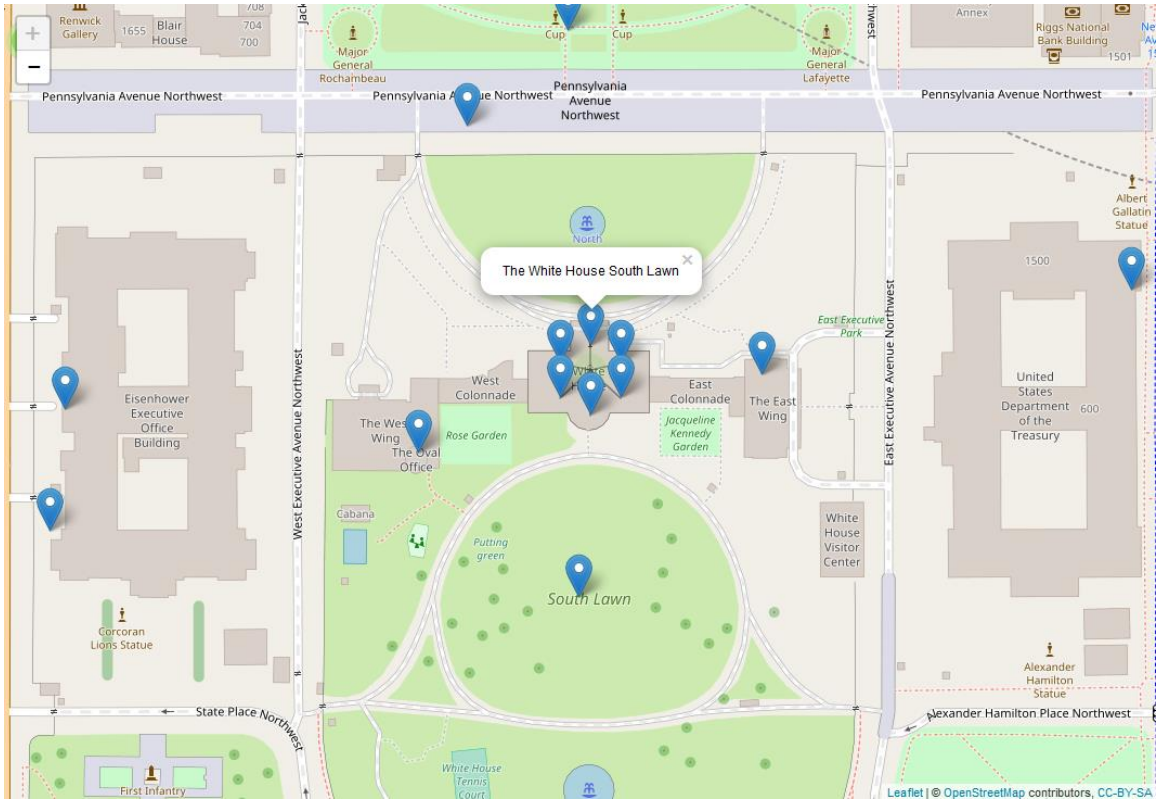

Figure 12. Zoomed in View of White House Instagram Locations

4.6 Summary

The applications of the data derived from this package are plentiful. With the many R packages at the disposal of users, visualization and analysis can be accomplished with style and ease. The data provided allows for media, user, and location driven analysis. As this package matures, these procedures can be added to the functionality of the package, but it currently serves an important role of providing useful and unique social media data to analysts in the intelligence community.

V. Conclusions and Recommendations

5.1 Overview

The potential for the `instaExtract` package and the practices this thesis hopes to encourage in the military analytical community has been shown to be promising, but the long-term success of this and future programs will rely on the adoption rate of users and the continued support of technical organization like the AFIT Data Science Lab. At its core, the `instaExtract` package achieve its operational goal of providing resources for analysts to extract Instagram data. Additionally, this thesis attempts to accomplish this in a manner with forward reaching implications. R is a heavily supported, open source language, with new resources created by the community on a routine basis. Choosing to host the package on GitHub allows the package the benefit of future support and user involvement.

5.2 The `instaExtract` Package

The `instaExtract` package boasts most of the same functionality offered by similar available programs while exhibiting even more paths for it to grow. First and foremost, the package achieves its goal of scraping Instagram for data. It offers search functions and a suite of get functions that allow a user to obtain almost any information that Instagram tracks. Furthermore, it has a collection of location mapping functions that mimic the features that were traditionally limited to programs interacting with the Instagram API. All of these functions can be operated without authorization and the need to have log-in information for an Instagram account.

The package also has many areas for future development. While most information can be retrieved without authorization, or otherwise inferred from available data, there

are some areas of Instagram that require log-in credentials. Future work should be done to ensure the robustness of operations, such as compilation and use time error and compatibility checking. As a final step towards professionalization, the package could also be published to the Comprehensive R Archive Network (CRAN). CRAN is, as its name suggests, a comprehensive archive of R packages that amount to the most commonly used and relied upon functions in R. Publishing a package to CRAN allows users to be sure of a certain set of requirements and accessibility that are enforced by the CRAN maintainers.

In addition to work that falls within the purview of good coding practices, the `instaExtract` package has many avenues of growth that would benefit users and increase its applications. While certainly not exhaustive, this thesis suggests a few ideas considered to be worthy of attention:

- Extending Location Mapping Functionality – Currently, the location mapping functions can only access those locations that are listed through the explore/locations page on Instagram. Adding functions that allow users to quickly add their choice of locations to a mapping or that seek out other locations on their own could further enhance the usefulness of this collection.
- A Shiny Application – Shiny is an R package that creates applications to allow users to interact with a user interface rather than command lines in R. A Shiny app would increase the reach of the package and potentially help inform its users of all of its functions
- Proxy Settings – The `instaExtract` package uses the `cURL` package to interface with internet. As such, it does not have well behaved proxy settings.

Incorporating better ways to control these settings could allow for a user to work through a proxy, integrating with a local network or masking the location of their requests.

- Other Scrapping Techniques –How the JSON link approach to data scrapping is susceptible to changes in Instagram's structuring has already been discussed. However, it would be possible for future developers to allow for users to scrape their selected data through other means. One such option is the Selenium package, which uses conventional browsers to navigate between webpages. Although it is exclusive, gaining access to the Instagram API could enable many new options and considerably accelerate existing ones.
- Analysis Functions – The package was designed around the goal of scrapping data, but it would be natural to include analytical functions in the future. The current package lends itself to sentiment analysis, anomaly detection, and more. With methods of pulling the images that are represented currently only as URLs would also allow for image recognition.

5.3 The AFIT Data Science Lab

A major academic pursuit of this paper is to validate the viability and incentives of programs hosted on a centralized and collaborative platform, such as the AFIT Data Science Lab. The GitHub platform will greatly simplify and facilitate the future development of this package. Users not only have access to a reliable and up-to-date version of the software, but can maintain a responsive dialog with the maintainers of the code as well as share the results of their individual efforts. GitHub's version control also

enables analysis with a focus on reproducibility by storing previous versions of the package.

The long-term value of these pursuits will be determined by the adoption rate of similar practices by other analytical software and the amount of support given to software development organizations. As discussed, packages stored in a centralized location receive a multiplicative benefit of enjoying a higher user base and more engaged developers. But of course, this requires the support structure to address development issues and to create new packages to meet user needs. Ensuring this support staff is valued and funded will require further commitment to similarly focused programs and research that can validate the worth of such programs.

5.4 Summary

It is clear that `instaExtract` package adds value to the analytical and intelligence communities, but this research acknowledges that there is more work to be done in solidifying the robustness of this package and validating and realizing a collaborative and comprehensive hub for analytic resources. Future research should be done into more data acquisition software and the statistical benefits of platforms like GitHub and their effects on DoD workflows.

Appendix A

Link to the instaExtract package: https://github.com/AFIT-R/instaExtract

Appendix B

id	name	slug	city_ID	city_Name	city_Slug	country_ID	country_Name	country_Slug	latitude	longitude
213480180	Washington, District of Columbia	washington-district-of-columbia	c2427178	Washington	washington-united-states	US	United States	united-states	38.8951	-77.0367
225931565	The Obama White House	the-obama-white-house	c2427178	Washington	washington-united-states	US	United States	united-states	38.89768	-77.03655
15712	Lincoln Memorial	lincoln-memorial	c2427178	Washington	washington-united-states	US	United States	united-states	38.889444	-77.050278
3001994	United States Capitol	united-states-capitol	c2427178	Washington	washington-united-states	US	United States	united-states	38.8897301	-77.0070362
214773851	Washington Monument National Monument	washington-monument-national-monument	c2427178	Washington	washington-united-states	US	United States	united-states	38.8890235	-77.0331092
4366681	National Gallery of Art	national-gallery-of-art	c2427178	Washington	washington-united-states	US	United States	united-states	38.8913397	-77.0196344
82474402	Smithsonian's National Zoo and Conservation Biology Institute	smithsonians-national-zoo-and-conservation-biology-institute	c2427178	Washington	washington-united-states	US	United States	united-states	38.9299778	-77.0511297
1191441824276880	The White House	the-white-house	c2427178	Washington	washington-united-states	US	United States	united-states	38.8968447	-77.0366049
235453813	Nationals Park	nationals-park	c2427178	Washington	washington-united-states	US	United States	united-states	38.8732565	-77.0075808
372247132	Capital One Arena	capital-one-arena	c2427178	Washington	washington-united-states	US	United States	united-states	38.89795	-77.02096
262515071	Smithsonian National Museum of African American History and Culture	smithsonian-national-museum-of-african-american-history-and-culture	c2427178	Washington	washington-united-states	US	United States	united-states	38.8911015	-77.0325639
373555	Renwick Gallery	renwick-gallery	c2427178	Washington	washington-united-states	US	United States	united-states	38.898867	-77.039447
236471522	The Mall (Washington DC)	the-mall-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.89	-77.0236111
214720506	The Capitol, Washington D.C.	the-capitol-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.8898395	-77.0154594
279173	Howard University	howard-university	c2427178	Washington	washington-united-states	US	United States	united-states	38.92314	-77.02158
214638086	Smithsonians National Museum of Natural History	smithsonians-national-museum-of-natural-history	c2427178	Washington	washington-united-states	US	United States	united-states	38.8910781	-77.026232
2222215	National Building Museum	national-building-museum	c2427178	Washington	washington-united-states	US	United States	united-states	38.89779	-77.01752
838999	National Museum of American History	national-museum-of-american-history	c2427178	Washington	washington-united-states	US	United States	united-states	38.8913741	-77.0299286
175770	Hirshhorn Museum and Sculpture Garden	hirshhorn-museum-and-sculpture-garden	c2427178	Washington	washington-united-states	US	United States	united-states	38.8883276	-77.0229156
212896512	National Air and Space Museum, Smithsonian Institution	national-air-and-space-museum-smithsonian-institution	c2427178	Washington	washington-united-states	US	United States	united-states	38.8881412	-77.0198422
42620	9:30 Club	930-club	c2427178	Washington	washington-united-states	US	United States	united-states	38.91803	-77.02363
214513963	Union Station, Washington D.C.	union-station-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.8975	-77.00621
2150339	The George Washington University	the-george-washington-university	c2427178	Washington	washington-united-states	US	United States	united-states	38.8982042	-77.050286
139456	Jefferson Memorial	jefferson-memorial	c2427178	Washington	washington-united-states	US	United States	united-states	38.881111	-77.036667
576233589	Lijst van bekende mensen uit Washington D.C.	lijst-van-bekende-mensen-uit-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.8951111	-77.0366667
216550363	National Portrait Gallery, Smithsonian Institution	national-portrait-gallery-smithsonian-institution	c2427178	Washington	washington-united-states	US	United States	united-states	38.8977641	-77.0229628
849479	The Library of Congress	the-library-of-congress	c2427178	Washington	washington-united-states	US	United States	united-states	38.8887863	-77.0058375
1009997177	Georgetown, DC	georgetown-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.9026909	-77.0494537
251483	Capitol Hill	capitol-hill	c2427178	Washington	washington-united-states	US	United States	united-states	38.889722	-77.011111

Appendix C

X	id	name	slug	city_ID	city_Name	city_Slug	country_ID	country_Name	country_Slug	latitude	longitude
1	213480180	Washington, District of Columbia	washington-district-of-columbia	c2427178	Washington	washington-united-states	US	United States	united-states	38.8951	-77.0367
2	225931565	The Obama White House	the-obama-white-house	c2427178	Washington	washington-united-states	US	United States	united-states	38.89768	-77.03655
3	15712	Lincoln Memorial	lincoln-memorial	c2427178	Washington	washington-united-states	US	United States	united-states	38.889444	-77.050278
5	214773851	Washington Monument National Monument	washington-monument-national-monument	c2427178	Washington	washington-united-states	US	United States	united-states	38.8890235	-77.0331092
8	1191441824276880	The White House	the-white-house	c2427178	Washington	washington-united-states	US	United States	united-states	38.89684467	-77.03660488
10	372247132	Capital One Arena	capital-one-arena	c2427178	Washington	washington-united-states	US	United States	united-states	38.89795	-77.02096
11	262515071	Smithsonian National Museum of African American History and Culture	smithsonian-national-museum-of-african-american-history-and-culture	c2427178	Washington	washington-united-states	US	United States	united-states	38.89110149	-77.03256389
12	373555	Renwick Gallery	renwick-gallery	c2427178	Washington	washington-united-states	US	United States	united-states	38.898867	-77.039447
13	236471522	The Mall (Washington DC)	the-mall-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.89	-77.02361111
16	214638086	Smithsonians National Museum of Natural History	smithsonians-national-museum-of-natural-history	c2427178	Washington	washington-united-states	US	United States	united-states	38.89107814	-77.02623198
18	838999	National Museum of American History	national-museum-of-american-history	c2427178	Washington	washington-united-states	US	United States	united-states	38.89137413	-77.02992863
19	175770	Hirshhorn Museum and Sculpture Garden	hirshhorn-museum-and-sculpture-garden	c2427178	Washington	washington-united-states	US	United States	united-states	38.88832764	-77.02291558
23	2150339	The George Washington University	the-george-washington-university	c2427178	Washington	washington-united-states	US	United States	united-states	38.8982042	-77.05028604
25	576233589	Lijst van bekende mensen uit Washington D.C.	lijst-van-bekende-mensen-uit-washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.89511111	-77.03666667
26	216550363	National Portrait Gallery, Smithsonian Institution	national-portrait-gallery-smithsonian-institution	c2427178	Washington	washington-united-states	US	United States	united-states	38.8977641	-77.02296276
28	1009997177	Georgetown, DC	georgetown-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.9026909	-77.0494537
30	2091823	National World War II Memorial	national-world-war-ii-memorial	c2427178	Washington	washington-united-states	US	United States	united-states	38.88935278	-77.04055556
31	26131	Newseum	newseum	c2427178	Washington	washington-united-states	US	United States	united-states	38.89313	-77.01935
33	590718213	Dupont Circle	dupont-circle	c2427178	Washington	washington-united-states	US	United States	united-states	38.9096	-77.0434
34	21614	Walter E. Washington Convention Center	walter-e-washington-convention-center	c2427178	Washington	washington-united-states	US	United States	united-states	38.9042406	-77.02325315
38	342980053	Tidal Basin	tidal-basin	c2427178	Washington	washington-united-states	US	United States	united-states	38.8837	-77.0389
39	215073	Reflecting Pool	reflecting-pool	c2427178	Washington	washington-united-states	US	United States	united-states	38.88926578	-77.04750981
40	683422	Madame Tussauds DC	madame-tussauds-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.89759	-77.0261699
41	75367954	Le Diplomate	le-diplomate	c2427178	Washington	washington-united-states	US	United States	united-states	38.91136386	-77.03157239
43	21783	International Spy Museum	international-spy-museum	c2427178	Washington	washington-united-states	US	United States	united-states	38.89699	-77.02329
44	218723464	National Museum of Women in the Arts	national-museum-of-women-in-the-arts	c2427178	Washington	washington-united-states	US	United States	united-states	38.90006	-77.02916
47	212897059	Washington, DC	washington-dc	c2427178	Washington	washington-united-states	US	United States	united-states	38.89511111	-77.03666667
53	1420252	Martin Luther King, Jr. National Memorial	martin-luther-king-jr-national-memorial	c2427178	Washington	washington-united-states	US	United States	united-states	38.88611111	-77.045
55	11195531	Smithsonians Freer and Sackler Galleries	smithsonians-freer-and-sackler-galleries	c2427178	Washington	washington-united-states	US	United States	united-states	38.88796	-77.02645

Appendix D

	comment_id		thumbnail_src	is_video	code	date	display_src	caption	video_v	dimensio	dimensio	edge_me	owner.id	comment	likes.cou	location_n	location_id
1	FALSE	17157344309893200	https://scontent-ort2	FALSE	BfPg1_oBi	1518751504	https://scontent-ort2-2.c	This is Radley. He is a very good	NA	1258	1080	0	209427535	0	0	0 Washingt	213480180
2	FALSE	17157344228655000	https://scontent-ort2	FALSE	BfPg14CFi	1518751503	https://scontent-ort2-2.c	Ahhhhh - one our very cozy Tile	NA	1080	1080	0	5599785536	0	0	0 The Obam	225931565
3	FALSE	17157344219732300	https://scontent-ort2	FALSE	BfPg13On	1518751503	https://scontent-ort2-2.c	Ugly cute ice cream--	NA	1349	1080	0	543733339	0	0	0 Lincoln M	15712
4	FALSE	17157343282717000	https://scontent-ort2	FALSE	BfPgdf9lS	1518751492	https://scontent-ort2-2.c	NA	NA	1253	1080	0	21769628	0	0	0 Washingt	214773851
5	FALSE	17157339150742600	https://scontent-ort2	FALSE	BfPgufJF2	1518751442	https://scontent-ort2-2.c	Natural light and film grain	NA	1080	1080	11	1170155535	0	11	The White	1191441824276880
6	FALSE	17157338490964400	https://scontent-ort2	FALSE	BfPgthsgc	1518751434	https://scontent-ort2-2.c	Here we are again.	NA	1350	1080	1	4689176897	0	1	Capital Or	372247132
7	FALSE	17157337752863500	https://scontent-ort2	FALSE	BfPgsc9FT	1518751426	https://scontent-ort2-2.c	stuffy caity from @the_shelter_	NA	1080	1080	2	1917442	0	2	Smithsoni	262515071
8	FALSE	17157336826170300	https://scontent-ort2	FALSE	BfPgrGpJ6	1518751415	https://scontent-ort2-2.c	<U+0410><U+0441><U+0435><U	NA	810	1080	0	4361163372	0	0	Renwick G	373555
9	FALSE	17157335490780200	https://scontent-ort2	FALSE	BfPgpKSB	1518751399	https://scontent-ort2-2.c	I could live in these fluffy sweet	NA	810	1080	0	244416823	0	0	The Mall (	236471522
10	FALSE	17157335285936900	https://scontent-ort2	FALSE	BfPgo3ND	1518751396	https://scontent-ort2-2.c	Check Us Out EVERY Saturday	NA	1080	1080	0	1919122075	0	0	Smithsoni	214638086
11	FALSE	17157333083675200	https://scontent-ort2	FALSE	BfPgIqQjV	1518751370	https://scontent-ort2-2.c	Check Us Out EVERY Saturday	NA	1080	1080	0	1919122075	0	0	National M	838999
12	FALSE	17157332817163900	https://scontent-ort2	FALSE	BfPgIRSBL	1518751367	https://scontent-ort2-2.c	Post workout lunch	NA	1080	1080	0	6239408513	0	0	Hirshhorn	175770
13	FALSE	17157331177291700	https://scontent-ort2	FALSE	BfPgi4jnu	1518751347	https://scontent-ort2-2.c	NO NEED TO TRY, IT COMES NAT	NA	1080	1080	1	440468706	0	1	The Georg	2150339
14	FALSE	17157299753965700	https://scontent-ort2	TRUE	BfPf1KCAi	1518751343	https://scontent-ort2-2.c	Had a great conversation about	4	421	750	4	2107074902	1	4	Lijst van b	576233589
15	FALSE	17157330526316700	https://scontent-ort2	FALSE	BfPgh77gs	1518751339	https://scontent-ort2-2.c	<f0><U+009F><U+0090><U+00BE	NA	1349	1080	2	3929073746	0	2	National F	216550363
16	FALSE	17157330440684800	https://scontent-ort2	FALSE	BfPgnz9H	1518751338	https://scontent-ort2-2.c	#LingerieParty <f0><U+009F><U	NA	1350	1080	3	1700378091	0	3	Georgetov	1009997177
17	FALSE	17157330067731000	https://scontent-ort2	FALSE	BfPghROH	1518751334	https://scontent-ort2-2.c	Travelling to Union Market	NA	1349	1080	0	710737204	0	0	National \	2091823
18	FALSE	17157326837425500	https://scontent-ort2	FALSE	BfPgdkYAI	1518751296	https://scontent-ort2-2.c	<U+30C8><U+3DEC><U+3DB8><U	NA	810	1080	8	4751747131	1	8	Newseum	26131
19	FALSE	17157323956696300	https://scontent-ort2	FALSE	BfPgYFfhI	1518751261	https://scontent-ort2-2.c	Had vaycay. Need runs. And	NA	1350	1080	3	6577297684	0	3	Dupont Ci	590718213
20	FALSE	17157319623396000	https://scontent-ort2	FALSE	BfPgSEhA	1518751210	https://scontent-ort2-2.c	It's not even that warm and I'm	NA	1080	1080	0	683549315	0	0	Walter E.	21614
21	FALSE	17157319267883800	https://scontent-ort2	FALSE	BfPgRjaBJ	1518751205	https://scontent-ort2-2.c	The #LBD gets a #bosslady	NA	1349	1080	8	3901997163	0	8	Tidal Basi	342980053
22	FALSE	17157317296814500	https://scontent-ort2	FALSE	BfPgOr1hI	1518751182	https://scontent-ort2-2.c	if you wanna come thru, there's	NA	1349	1080	11	22369224	0	11	Reflecting	215073
23	FALSE	17157316258296000	https://scontent-ort2	FALSE	BfPgNNrlv	1518751170	https://scontent-ort2-2.c	Here w/ Fashion designer/stylis	NA	1080	1080	3	2260503050	0	3	Madame]	683422
24	FALSE	17157312551377900	https://scontent-ort2	FALSE	BfPgHx4h	1518751125	https://scontent-ort2-2.c	A new DC mural, spied it today!	NA	1080	1080	0	535139382	0	0	Le Diplom	75367954
25	FALSE	17157311770601000	https://scontent-ort2	FALSE	BfPgGpHg	1518751116	https://scontent-ort2-2.c	NA	NA	1079	1080	1	5110557	0	1	Internatio	21783
26	FALSE	17157307729537100	https://scontent-ort2	FALSE	BfPgAw0h	1518751068	https://scontent-ort2-2.c	NOT TODAY, SATAN.	NA	1080	1080	15	3228039039	0	15	National M	218723464
27	FALSE	17157307182846500	https://scontent-ort2	FALSE	BfPf_95lS	1518751061	https://scontent-ort2-2.c	Not a good run, but a pretty nigh	NA	810	1080	3	640714	0	3	Washingt	212897059
28	FALSE	17157305220506000	https://scontent-ort2	FALSE	BfPf9HJHz	1518751038	https://scontent-ort2-2.c	Follow me at @corinnethedogm	NA	1080	1080	7	513837466	1	7	Martin Lut	1420252
29	FALSE	17157304194328000	https://scontent-ort2	FALSE	BfPf7nknS	1518751026	https://scontent-ort2-2.c	Ruth E. Carter	NA	719	1080	8	930631721	0	8	Smithsoni	11195531
30	FALSE	17157302065531100	https://scontent-ort2	FALSE	BfPf4hUA	1518751000	https://scontent-ort2-2.c	This is our children America. The	NA	1080	1080	1	2095237453	0	1	Vashingto	1024272801
31	FALSE	17157301527489600	https://scontent-ort2	FALSE	BfPf3vNB	1518750994	https://scontent-ort2-2.c	Conservatives will never champ	NA	1080	1080	10	4588361131	0	10	W Washin	133459
32	FALSE	17157299751199100	https://scontent-ort2	FALSE	BfPf1Jxgy	1518750973	https://scontent-ort2-2.c	Activo Con el #FlowLePew<U+2(	NA	480	480	2	281146026	0	2	Founding	295793
33	FALSE	17157298127165500	https://scontent-ort2	FALSE	BfPfyyhhl	1518750953	https://scontent-ort2-2.c	Roman vibes throwback	NA	1080	1080	10	213383616	2	10	CityCente	940940154
34	FALSE	17157295691030500	https://scontent-ort2	FALSE	BfPfvPpBJ	1518750924	https://scontent-ort2-2.c	boots with the fur	NA	1350	1080	48	20888638	3	48	Vietnam \	133162
35	FALSE	17157295217072200	https://scontent-ort2	FALSE	BfPfujgAp	1518750919	https://scontent-ort2-2.c	Leo has made a new best friend	NA	1297	1080	7	1497688573	1	7	DAR Const	565861
36	FALSE	17157294795390000	https://scontent-ort2	FALSE	BfPft8Olo	1518750914	https://scontent-ort2-2.c	"Dame un 'e poco love with	NA	719	1080	20	1527693133	0	20	The Hamil	7025540
37	FALSE	17157294476360800	https://scontent-ort2	FALSE	BfPfteh8r	1518750910	https://scontent-ort2-2.c	Had a lovely evening with my bi	NA	1080	1080	0	32220304	0	0	United Sta	215230297
38	FALSE	17157293336787000	https://scontent-ort2	FALSE	BfPfr0GHI	1518750896	https://scontent-ort2-2.c	I'm here, in my place, finally <U	NA	1080	1080	2	202686924	0	2	World Wa	464496570
39	FALSE	17157289822623100	https://scontent-ort2	FALSE	BfPfmtGlr	1518750854	https://scontent-ort2-2.c	Dont mean to break your heart t	NA	1080	1080	12	5974320067	1	12	Warner Th	84405
40	FALSE	17157286554426100	https://scontent-ort2	FALSE	BfPfh8unz	1518750815	https://scontent-ort2-2.c	NA	NA	1350	1080	23	3142988762	0	23	Smithsoni	213933592
41	FALSE	17157286025256500	https://scontent-ort2	FALSE	BfPfhLchh	1518750809	https://scontent-ort2-2.c	Sweet	NA	1350	1080	280	14232494	9	280	National (	265107589
42	FALSE	17157285000381000	https://scontent-ort2	FALSE	BfPffr_h5I	1518750797	https://scontent-ort2-2.c	Who wants to get crazy with me	NA	1162	1080	2	40285465	3	2	Washingt	293984628
43	FALSE	17157284623865700	https://scontent-ort2	FALSE	BfPffi7lDj	1518750792	https://scontent-ort2-2.c	#throwbackthursday Back in the	NA	1080	1080	7	47399013	2	7	Abraham	212895774
44	FALSE	17157283565883900	https://scontent-ort2	FALSE	BfPfdmZh	1518750780	https://scontent-ort2-2.c	With yo Natural hair, your soft s	NA	1350	1080	31	33708514	0	31	Awesome	85622440
45	FALSE	17157281872900800	https://scontent-ort2	FALSE	BfPfbIujP	1518750759	https://scontent-ort2-2.c	Not your average or typical <fO>	NA	1307	1080	40	51785802	1	40	The Park a	439392
46	FALSE	17157273469438600	https://scontent-ort2	TRUE	BfPfO6GA	1518750694	https://scontent-ort2-2.c	More Odd Louise on the rise #th	15	640	640	2	27368448	0	2	Trump Int	1024936088
47	FALSE	17157266903591300	https://scontent-ort2	TRUE	BfPfFWmj	1518750663	https://scontent-ort2-2.c	shhhh... i'm hiding <fO><U+009F	18	936	750	13	5961776691	0	13	Old Ebbitt	49111800
48	FALSE	17157260108160600	https://scontent-ort2	TRUE	BfPe7duIx	1518750661	https://scontent-ort2-2.c	...Flexin'... *Coming Soon*	30	750	750	15	2132504969	3	15	Washingt	217010719
49	FALSE	17157273168049200	https://scontent-ort2	FALSE	BfPfQeBIz	1518750656	https://scontent-ort2-2.c	Thnx Dalpino.	NA	809	1080	30	3259870568	1	30	Martin Lut	220416539
50	FALSE	17157270645006000	https://scontent-ort2	FALSE	BfPfKzDFf	1518750628	https://scontent-ort2-2.c	#TBT <fO><U+009F><U+0092><U+	NA	1080	1080	14	174509918	0	14	World Bar	81794
51	FALSE	17156896651837000	https://scontent-ort2	FALSE	BfPWqkO	1518746167	https://scontent-ort2-2.c	Meeting the President in Portla	NA	1080	1080	25493	4303258197	197	25493	Washingt	213480180
52	FALSE	17156789529905700	https://scontent-ort2	FALSE	BfPUOtnI	1518744890	https://scontent-ort2-2.c	18: Can I buy beer?	NA	1080	1080	749	4933061103	29	749	The Obam	225931565
53	FALSE	17156700431222600	https://scontent-ort2	FALSE	BfPSNBwI	1518743828	https://scontent-ort2-2.c	#TBT #POTUS44 @barackobama	NA	1080	1080	37	5179681	2	37	Lincoln M	15712
54	FALSE	17156676532331000	https://scontent-ort2	FALSE	BfPRqP_A	1518743543	https://scontent-ort2-2.c	TBT to When the House was a	NA	1349	1080	54	28897532	1	54	Washingt	214773851
55	FALSE	17156471944657600	https://scontent-ort2	FALSE	BfPNAi5B	1518741104	https://scontent-ort2-2.c	#americanmade #getittogether	NA	640	640	28	44798710	0	28	The White	1191441824276880
56	FALSE	17156174281806000	https://scontent-ort2	FALSE	BfPGPYRn	1518737556	https://scontent-ort2-2.c	Fuck a glass of wine, let's punish	NA	845	1080	100	222906177	11	100	Capital Or	372247132
57	FALSE	17155873201814400	https://scontent-ort2	FALSE	BfO_ZQAI	1518733967	https://scontent-ort2-2.c	Find me on Tinder	NA	1080	1080	299	1122797437	33	299	Smithsoni	262515071
58	FALSE	17155820973676300	https://scontent-ort2	FALSE	BfO-NP4n	1518733344	https://scontent-ort2-2.c	SPECIAL INSPIRATION EDITION:	NA	717	1080	8	194392055	0	8	Renwick G	373555
59	FALSE	17154402271654800	https://scontent-ort2	FALSE	BfOd8w9r	1518716432	https://scontent-ort2-2.c	celery <f0><U+009F><U+0091><(	NA	750	750	25	188238192	2	25	The Mall (	236471522
60	FALSE	17154400815226200	https://scontent-ort2	FALSE	BfOd6pUr	1518716415	https://scontent-ort2-2.c	Do something today that your	NA	1080	1080	21	32160014	0	21	Smithsoni	214638086

Appendix E

media_preview	thumbnail_src	is_video	code	date	display_src	caption	video_views	dimensions .height	dimensions .width	edge_medi a_preview_ like.count	owner.id	comment s.count	likes.co unt	gating_in fo.descri ption	gating_in fo.gating _type	gating_in fo.title
ACoFyoLlkxz0/L8q3l	https://scontent-ort2-2.cd	FALSE	Bfl9J0Jn18a	1518632129	https://scontent-	A member of the 83rd	NA	804	1080	22983	1.97E+08	60	22983	NA	NA	NA
ACobzlmxVtXzVAO/	https://scontent-ort2-2.cd	FALSE	Bf1q3dZn7mG	1518555432	https://scontent-	An F-22 Raptor fighter jet	NA	694	1080	22445	1.97E+08	65	22445	NA	NA	NA
ACoXywWEQK5zk9E/	https://scontent-ort2-2.cd	FALSE	BfG-Gd5HPGz	1518464853	https://scontent-	An Indiana Air National	NA	607	1080	22609	1.97E+08	60	22609	NA	NA	NA
ACocyfgZh8u1+f45h/	https://scontent-ort2-2.cd	FALSE	Be_ETmvnB4Q	1518199671	https://scontent-	#ICYMI: U.S. Air Force	NA	720	1080	17378	1.97E+08	52	17378	NA	NA	NA
ACoWvwIMvevBXP/	https://scontent-ort2-2.cd	FALSE	Be8kpU5HUYw	1518115963	https://scontent-	Senior Airman Adam	NA	576	1080	20090	1.97E+08	70	20090	NA	NA	NA
ACobzj8gLAnjtL/z+d/	https://scontent-ort2-2.cd	FALSE	Be56xZxnJnt	1518026900	https://scontent-	Airman 1st Class Colin	NA	697	1080	11094	1.97E+08	25	11094	NA	NA	NA
ACocFZZACv8A9arV/	https://scontent-ort2-2.cd	FALSE	Be3U0YsHxOv	1517939893	https://scontent-	Airman 1st Class Kaylon	NA	721	1080	23499	1.97E+08	107	23499	NA	NA	NA
ACoc3aWkopiAnFv/	https://scontent-ort2-2.cd	FALSE	BeOwcCQnAE7	1517853710	https://scontent-	Airmen assigned to the	NA	721	1080	19023	1.97E+08	38	19023	NA	NA	NA
ACoqsgc1PuYHjk+m/	https://scontent-ort2-2.cd	TRUE	BewiRjxn116	1517712121	https://scontent-	#Repost	76382	640	640	9254	1.97E+08	87	9254	NA	NA	NA
ACohFiTFgn5ngHjU/	https://scontent-ort2-2.cd	FALSE	Bevw1fgHt8y	1517686146	https://scontent-	#Repost	NA	864	1080	17603	1.97E+08	39	17603	NA	NA	NA
ACoc5wCnb5Tgda8l/	https://scontent-ort2-2.cd	FALSE	BetAm52n3tC	1517593751	https://scontent-	Airmen assigned to the	NA	726	1080	18188	1.97E+08	48	18188	NA	NA	NA
ACohuW8zM7ByNu/	https://scontent-ort2-2.cd	FALSE	BeqGdC2n-9o	1517496154	https://scontent-	Developing International	NA	864	1080	14097	1.97E+08	27	14097	NA	NA	NA
ACobQgIZV5fGP5Yz/	https://scontent-ort2-2.cd	FALSE	Ben77mVHR0U	1517423528	https://scontent-	Staff Sgt. Michael	NA	713	1080	21184	1.97E+08	89	21184	NA	NA	NA
ACocjSKrcKI52fcf+6e/	https://scontent-ort2-2.cd	FALSE	BeIXtrGHIDv	1517337431	https://scontent-	Tech Sgt. Matthew	NA	721	1080	15559	1.97E+08	37	15559	NA	NA	NA
ACoX5mnhaEGa0bV/	https://scontent-ort2-2.cd	FALSE	BeiyWwfnVwQ	1517250736	https://scontent-	An Atlas V rocket carrying	NA	607	1080	19678	1.97E+08	52	19678	NA	NA	NA
ACoXqC1l6kY/EU/y/	https://scontent-ort2-2.cd	FALSE	Bea_T5UnvWm	1516989087	https://scontent-	A student with	NA	606	1080	22522	1.97E+08	68	22522	NA	NA	NA
ACocxtEUEAEkEHGO/	https://scontent-ort2-2.cd	FALSE	BeWbI-HnFr	1516900027	https://scontent-	U.S. Air Force Staff Sgt.	NA	721	1080	15703	1.97E+08	33	15703	NA	NA	NA
ACoVQg2RnrSBwQv/	https://scontent-ort2-2.cd	FALSE	BeV3SU4H7La	1516817113	https://scontent-	U.S. Air Force F-15C Eagle	NA	565	1080	21890	1.97E+08	63	21890	NA	NA	NA
ACob2tvpzUT57PvK/	https://scontent-ort2-2.cd	FALSE	BeTRzVjnus9	1516730352	https://scontent-	Members of the 146th Air	NA	718	1080	29747	1.97E+08	112	29747	NA	NA	NA
ACog6IEZxnKViX2Ip/	https://scontent-ort2-2.cd	FALSE	BeQuT-8HfpC	1516644636	https://scontent-	Capt. Zoe Kotnik, 55th	NA	846	1080	41541	1.97E+08	286	41541	NA	NA	NA
ACob2VkO43Ike6g/	https://scontent-ort2-2.cd	FALSE	Be6h-QqnY_C	1516382601	https://scontent-	A U.S. Air Force F-15	NA	718	1080	24327	1.97E+08	67	24327	NA	NA	NA
ACob3AQwyDkeO5j/	https://scontent-ort2-2.cd	FALSE	BeGeWntHymU	1516300725	https://scontent-	Paratroopers assigned to	NA	718	1080	19947	1.97E+08	66	19947	NA	NA	NA
ACoV2hmSL7x4HX2/	https://scontent-ort2-2.cd	FALSE	BeD6eRIn3JA	1516214804	https://scontent-	A 153rd Airlift Wing	NA	565	1080	20979	1.97E+08	85	20979	NA	NA	NA
ACoXxY3OPcVbiXzR/	https://scontent-ort2-2.cd	FALSE	BeAyL86n6Ad	1516109796	https://scontent-	YEAR IN PHOTOS IS	NA	424	750	10355	1.97E+08	25	10355	NA	NA	NA
ACobjHuDj1FT4Ce/	https://scontent-ort2-2.cd	FALSE	Bd2-Z9hHrIN	1515780658	https://scontent-	Capt. Michael Swartz,	NA	718	1080	16383	1.97E+08	61	16383	NA	NA	NA
ACoWpDnrTg3IHpj/	https://scontent-ort2-2.cd	FALSE	Bd0YLR2nkX4	1515693506	https://scontent-	Two A-10 Thunderbolt IIs	NA	579	1080	24892	1.97E+08	95	24892	NA	NA	NA
ACobxlXF5gU0V/pw	https://scontent-ort2-2.cd	FALSE	Bdx1La5nv1r	1515608048	https://scontent-	Senior Airman Avery	NA	719	1080	30813	1.97E+08	175	30813	NA	NA	NA
ACoXQQDQpA8cfVT/	https://scontent-ort2-2.cd	FALSE	BdvOBByHa5b	1515520407	https://scontent-	Hill Air Force Base F-35A	NA	608	1080	27976	1.97E+08	120	27976	NA	NA	NA
ACoC57FifQIpQmet/	https://scontent-ort2-2.cd	FALSE	BdsgCon0pj	1515435727	https://scontent-	A U.S. Air Force	NA	720	1080	18625	1.97E+08	40	18625	NA	NA	NA
ACod5OAtgDkmIaJf/	https://scontent-ort2-2.cd	FALSE	BdIOYoTnU5i	1515185056	https://scontent-	Airman 1st Class Jorden	NA	771	1080	14591	1.97E+08	45	14591	NA	NA	NA
NA	https://scontent-ort2-2.cd	FALSE	BdiHpU9HzMZ	1515080860	https://scontent-	The governor dedared as t	NA	500	750	36611	1.97E+08	212	36611	NA	NA	NA
ACoc5vFAFaltO5yDj/	https://scontent-ort2-2.cd	FALSE	Bdfuca_HCA7	1515000538	https://scontent-	A U.S. Air Force crew	NA	720	1080	15389	1.97E+08	38	15389	NA	NA	NA
ACobfMdx2jsK25fy/	https://scontent-ort2-2.cd	FALSE	BddOWRdnVqK	1514916602	https://scontent-	Elinor Otto, better known	NA	718	1080	19307	1.97E+08	48	19307	NA	NA	NA
ACoY57FLipTHhN+a/	https://scontent-ort2-2.cd	FALSE	Bd52lh1noG9	1514568599	https://scontent-	An Airman assigned to	NA	617	1080	19447	1.97E+08	61	19447	NA	NA	NA
ACocyZbiSY5ck+g7D/	https://scontent-ort2-2.cd	FALSE	BdQ5JdYn6f3	1514482386	https://scontent-	Staff Sgt. Richard Murkin,	NA	720	1080	20520	1.97E+08	55	20520	NA	NA	NA
ACoZmvtNOTJAPco/	https://scontent-ort2-2.cd	FALSE	BdNsGoNHOCt	1514395331	https://scontent-	Airmen from Nellis and	NA	644	1080	16236	1.97E+08	65	16236	NA	NA	NA
ACoenutQTIFJDD+6/	https://scontent-ort2-2.cd	FALSE	Bd8Ka9HHJA_	1513975019	https://scontent-	Airmen fold the flag durin	NA	780	1080	26104	1.97E+08	138	26104	NA	NA	NA
ACoc5zFGK2p0kwR/	https://scontent-ort2-2.cd	FALSE	Bc-Pm5FH8pI	1513877075	https://scontent-	A C-130J Super Hercules	NA	720	1080	20631	1.97E+08	74	20631	NA	NA	NA
ACocDFfGa1vWjG6u/	https://scontent-ort2-2.cd	FALSE	Bc7vICDn3rj	1513792936	https://scontent-	A Republic of Singapore	NA	720	1080	15379	1.97E+08	50	15379	NA	NA	NA
ACoYtBakCUy5Sjiba/	https://scontent-ort2-2.cd	FALSE	Bc5KSnrHoX7	1513706515	https://scontent-	A C-130J Super Hercules	NA	621	1080	20057	1.97E+08	37	20057	NA	NA	NA
ACoaxaKKK9E5woo/	https://scontent-ort2-2.cd	FALSE	Bc2nH3H0cS	1513620966	https://scontent-	A B-1B Lancer from	NA	685	1080	18784	1.97E+08	55	18784	NA	NA	NA
ACoXsf2T89SD/w89/	https://scontent-ort2-2.cd	TRUE	BcupjoZHkoY	1513354495	https://scontent-	C-130J aircraft from the	54602	607	1080	14771	1.97E+08	74	14771	NA	NA	NA
ACofOVMUcbxx1pXe/	https://scontent-ort2-2.cd	FALSE	BcsDMIaH3zx	1513266588	https://scontent-	General Joseph L. Lengyel	NA	810	1080	13653	1.97E+08	45	13653	NA	NA	NA
ACoboRhcAFckeh/X/	https://scontent-ort2-2.cd	FALSE	BcpxDuQHEiH	1513189969	https://scontent-	Airmen from the 820th	NA	718	1080	20432	1.97E+08	74	20432	NA	NA	NA
ACob1hIu/wCVSAS/	https://scontent-ort2-2.cd	FALSE	Bcnlh-vgtuX	1513101612	https://scontent-	A special missions	NA	717	1080	16966	1.97E+08	25	16966	NA	NA	NA
ACocwVV5dp4k3/x/	https://scontent-ort2-2.cd	FALSE	BcknOSTHueo	1513017041	https://scontent-	A New Jersey Air	NA	720	1080	21497	1.97E+08	51	21497	NA	NA	NA
ACobga1SfomwIihb/	https://scontent-ort2-2.cd	FALSE	Bcc1zOYnArp	1512756248	https://scontent-	A U.S. Air Force A-10	NA	718	1080	23479	1.97E+08	100	23479	NA	NA	NA
ACobDVFSVkjUsA/u/	https://scontent-ort2-2.cd	FALSE	BcaTAO8n1Jw	1512670895	https://scontent-	A U-2 Dragon Lady lands	NA	714	1080	26887	1.97E+08	125	26887	NA	NA	NA
ACocStVLHA5J4Fay/	https://scontent-ort2-2.cd	FALSE	BcXtYz9HLTm	1512584065	https://scontent-	Members of the	NA	720	1080	16932	1.97E+08	63	16932	NA	NA	NA
ACocQoX4kg/UwC/	https://scontent-ort2-2.cd	FALSE	BcVHEw1ng8u	1512496869	https://scontent-	Airmen from the 26th	NA	720	1080	16524	1.97E+08	29	16524	NA	NA	NA
NA	https://scontent-ort2-2.cd	FALSE	BcSleFmnnyZ	1512412142	https://scontent-	432nd Aircraft	NA	386	640	13993	1.97E+08	20	13993	NA	NA	NA
ACodGQsAMnoKiiu/	https://scontent-ort2-2.cd	FALSE	Bc68hIqnm3J	1512155790	https://scontent-	Secretary of the Air Force	NA	765	1080	24917	1.97E+08	198	24917	NA	NA	NA
ACoa047dU9Fzp7QK/	https://scontent-ort2-2.cd	FALSE	Bcifnc4HXjx	1512073528	https://scontent-	U.S. Air Force Staff Sgt.	NA	673	1080	16549	1.97E+08	52	16549	NA	NA	NA
ACofIsofLjzjB71HFZ/	https://scontent-ort2-2.cd	FALSE	BcFUthnFJW	1511980052	https://scontent-	Firefighters with the 1st	NA	810	1080	19033	1.97E+08	52	19033	NA	NA	NA
ACocxYUzXQWqbRX/	https://scontent-ort2-2.cd	FALSE	BcDQE8YHVXX	1511897609	https://scontent-	An Airman from the 92nd	NA	720	1080	14289	1.97E+08	30	14289	NA	NA	NA
ACokSuIxQDjuQuC/	https://scontent-ort2-2.cd	FALSE	BcAnne2nare	1511809288	https://scontent-	A KC-10 Extender from	NA	556	640	23612	1.97E+08	49	23612	NA	NA	NA
ACoaYoqZCZqOo8M/	https://scontent-ort2-2.cd	FALSE	BbzXdaoHKiN	1511364609	https://scontent-	A U.S. Air Force HH-60	NA	683	1080	22467	1.97E+08	62	22467	NA	NA	NA
ACocu55pPf2npVMC/	https://scontent-ort2-2.cd	FALSE	BbxDrKYHdNN	1511287127	https://scontent-	A U.S. Air Force C-5M	NA	720	1080	25173	1.97E+08	85	25173	NA	NA	NA
ACoc59R54raFprHm/	https://scontent-ort2-2.cd	FALSE	BbufFASnuoe	1511201044	https://scontent-	U.S. Air Force Capt. James	NA	721	1080	22819	1.97E+08	49	22819	NA	NA	NA

Appendix F

id	text	created_at	owner.id	owner.profile_	owner.username
17902716826081000	Nice!!	1516644761	1257719479	https://sconte	soph_ontwowheels
17892879493145000	<f0><U+009F><U	1516644775	306180651	https://sconte	sallykoonin_123
17920219525007200	Good girl! <U+27	1516644783	31383965	https://instagr	cc_the_dj
17846245039238900	ah	1516644792	1581502399	https://sconte	big_candycane_lane_madeosteel
17929182385004300	Get it.	1516644802	3016165931	https://sconte	dennisgivens
17859767317219600	<f0><U+009F><U	1516644827	3922217374	https://sconte	sam.deylami4054
17919995560054000	its a timex expe	1516644852	1581502399	https://sconte	big_candycane_lane_madeosteel
17902404370091000	@anna.ekstroml	1516644867	2222529160	https://sconte	ville360
17846218849241100	Nice	1516644876	5713619573	https://sconte	airforce.ir
17846242351233800	That is awesome	1516644878	270062153	https://sconte	rozyp77
17892406873161800	Eyebrows on fle	1516645059	33344140	https://sconte	treybrah_
17907310702129400	<f0><U+009F><U	1516645111	39121918	https://sconte	mekala333
17896066945138500	Shaw AFB repres	1516645119	34501017	https://sconte	chrisastro
17923311298058600	SALUTE!!	1516645150	6481124	https://sconte	paulagoble
17919239926045500	@idcvickyy u car	1516645242	415630583	https://sconte	joha_balay
17893049878150400	YEA! Viper pilots	1516645267	3266740512	https://sconte	onyxreaper
17902779208084300	<f0><U+009F><U	1516645297	210947008	https://sconte	crazy_gpigz
17846326054239800	Ville, jag är 14<f(	1516645301	857171570	https://sconte	anna.ekstromh
17907677278112200	#goals<U+2764><	1516645343	22737475	https://sconte	giaxlana
17895607483134500	<f0><U+009F><U	1516645352	4608014410	https://sconte	ciopaer_to
17919464815010700	@amanduh0429	1516645382	400073675	https://sconte	cacraig11
17879493088196800	<f0><U+009F><U	1516645436	4979572474	https://sconte	jaymon3002
17919810205032800	<f0><U+009F><U	1516645436	4979572474	https://sconte	jaymon3002
17894531890188700	Sickkkk	1516645519	1981970767	https://sconte	just.__tin___can_
17920278703051200	<U+2764><U+FE0	1516645582	285396048	https://sconte	j_yvette_l
17894926300130200	Can you get her	1516645687	295804343	https://sconte	lo_lights_
17902481515084500	NND PBSBHBJBL	1516645713	6330870614	https://sconte	noa727
17906989870099700	Veeeeery niiiiiii	1516645731	3579379587	https://sconte	koroush_imani
17879685745196500	@lo_lights_ I ask	1516645755	424550255	https://sconte	mo_elizabeth3
17920758868017800	@frankpaint	1516645776	46654786	https://sconte	laurenpainter_
17915043367070100	@mo_elizabeth3	1516645806	295804343	https://sconte	lo_lights_
17924786815057300	Found <f0><U+0(	1516645911	2266396868	https://sconte	a.badhisutawan
17921940580013500	<f0><U+009F><U	1516645921	3514876709	https://sconte	fawzinelfarr
17907142621097000	Bravo Bravo Brav	1516645957	4755191572	https://sconte	miro.jaume
17893056922148200	Great picture , th	1516646062	4699730528	https://sconte	lmma2003
17904741163126700	You shot the @u	1516646228	3228699880	https://sconte	ian_t_wilson_
17905736824126400	Oh wait wrong p	1516646264	3228699880	https://sconte	ian_t_wilson_
17905311511124800	#CooL	1516646416	4285556838	https://sconte	marlilny
17860723798209200	@jadentkorth Eg	1516646421	2064800284	https://sconte	anthonyliz_
17920338988032700	@anthonyliz_ HU	1516646464	246687008	https://sconte	jadentkorth
17893257619192700	BEAUTY AND STR	1516646513	4871254374	https://sconte	jjascosmos
17920623730011200	@jamaican_johr	1516646584	487458548	https://sconte	cyrbuiltsedan
17892821923152100	YGBSM	1516646811	31484972	https://sconte	isthisguyserious
17895336844131900	Beauty with the	1516646905	5880550922	https://sconte	suman_ayush13
17919505357015800	Wow strong figh	1516646962	2039822000	https://sconte	ahmedn22n
17918963671061900	Fire	1516646979	2039822000	https://sconte	ahmedn22n
17921023285040300	Too cool! @morg	1516647093	2017650474	https://sconte	adrianna_morgart
17860535284208300	<f0><U+009F><U	1516647293	191343376	https://sconte	pepegrillo89
17907340705106600	@ihashubeita	1516647348	2345937556	https://sconte	atilio_vurvopolos
17911435531078600	LOVE this!!	1516647427	27753143	https://sconte	jacqgould
17920140301030800	A dream that i w	1516647437	3245473458	https://sconte	call_me_numpt
17910967678076700	@charlie_collie	1516647440	15755357	https://sconte	passion4nyc
17920419412009400	Does she have Ir	1516647462	5527825192	https://sconte	araujo253
17905484125127500	Very good	1516647881	6796434290	https://sconte	aminallah.a
17889439036173600	#TeamShaw	1516647902	365607661	https://sconte	beningold318
17861953780206500	@anna.ekstroml	1516648016	2222529160	https://sconte	ville360
17902468441091800	<f0><U+009F><U	1516648216	5910717791	https://sconte	shocktherapy.jn
17895479689186600	<f0><U+009F><U	1516648412	1370345057	https://sconte	joeypr69
17907772753102700	BIUTIFUL	1516648554	2870647697	https://sconte	shahin_e_aseman
17893612492192200	@jonathan_viv1	1516648563	1184245960	https://sconte	jmonje14

Bibliography

Aggarwal, C. (2011). *Social Network Data Analytics*. *Vasa*. https://doi.org/10.1007/978-1-4419-8462-3

Aggarwal, C. & Yu, P. S. (2005). Online Analysis of Community Evolution in Data Streams. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 56–67. https://doi.org/10.1137/1.9781611972757.6

Andriopoulou, F. & Lymberopoulos, D. K. (2012). Artificial Intelligence Applications and Innovations. *IFIP Advances in Information and Communication Technology* (Vol. 382). https://doi.org/10.1007/978-3-642-33412-2

Arcega, R. (2018). instagram-scraper. Retrieved from https://github.com/rarcega/instagram-scraper

Ashraf, S. S., Verma, S., & Tech, M. (2016). A Survey on Sentiment Analysis Techniques on Social Media Data, *3*(3), 65–68. Retrieved from http://epubs.siam.org/doi/abs/10.1137/1.9781611972757.6

Barberá, P. (2017). instaR. Retrieved from https://github.com/pablobarbera/instaR

Bian, J., Yoshigoe, K., Hicks, A., Yuan, J., He, Z., Xie, M., Guo, Y., Prosperi, M., Salloum, R., Modave, F. (2016). Mining Twitter to Assess the Public Perception of the "Internet of Things". *PLoS ONE*, *11*(7), 1–14. Retrieved from http://10.0.5.91/journal.pone.0158450

Chan, C. K., Vasardani, M., & Winter, S. (2014). Leveraging Twitter to Detect Event Names Associated with a Place. *Journal of Spatial Science*, *59*(1), 137–155. Retrieved from http://10.0.4.56/14498596.2014.852073

Chaykowski, K. (2016). Instagram's Big Picture. *Forbes*, *198*(2), 62–69. Retrieved from http://widgets.ebscohost.com/prod/customerspecific/ns000290/authentication/index.php?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib,uid&db=heh&AN=117050301&lang=pt-br&site=eds-live&scope=site%5Cnhttp://content.ebscohost.com

Cheng, J. (2017). leaflet. Retrieved from https://cran.r-project.org/web/packages/leaflet/leaflet.pdf

Erlandsson, F., Bródka, P., Borg, A., & Johnson, H. (2016). Finding Influential Users in Social Media Using Association Rule Learning. *Entropy*, *18*(5), 1–15. Retrieved from http://10.0.13.62/e18050164

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola,

F. (2013). Web Scraping Technologies in an API World. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026

Hello World - GitHub Guides. (n.d.). Retrieved February 3, 2018, from https://guides.github.com/activities/hello-world/

Injadat, M., Salo, F. & Nassif, A. B. (2016). Data Mining Techniques in Social Media: A Survey. *Neurocomputing*, *214*, 654–670. Retrieved from http://10.0.3.248/j.neucom.2016.06.045

Instagram Scraper - Im Risto - Internet Marketing Blog. (n.d.). Retrieved February 13, 2018, from http://imristo.com/social-media-tools/instagram-scraper/

Jang, S. & Lee, E. (2009). Reliable Mobile Application Modeling Based on Open API. In D. Śl\kezak, T. Kim, A. Kiumi, T. Jiang, J. Verner, & S. Abrahão (Eds.), *Advances in Software Engineering* (pp. 168–175). Berlin, Heidelberg: Springer Berlin Heidelberg.

JSON. (n.d.). Retrieved February 13, 2018, from https://www.json.org/

Leese, C. B. (2015). Mining Social Media for Intel. *U.S. Naval Institute Proceedings*, *141*(8), 82–83. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=109023176&site=ehost-live

Li, H., Zha, Z.-J., Huet, B., & Tian, Q. (2016). Guest Editorial: Large-Scale Multimedia Content Analysis on Social Media. *Multimedia Tools and Applications*, *75*(3), 1365–1369. https://doi.org/10.1007/s11042-016-3255-z

Loftus, E. F. (2011). Intelligence Gathering Post-9/11. *American Psychologist*, *66*(6), 532–541. Retrieved from http://10.0.4.13/a0024614

Lv, J. & Guo, J. (2016). Mining Communities in Social Network Based on Information Diffusion. *IEEJ Transactions on Electrical & Electronic Engineering*, *11*(5), 604–617. Retrieved from http://10.0.3.234/tee.22278

Mezghani, M., Péninou, A., Zayani, C. A., Amous, I., & Sèdes, F. (2017). Producing Relevant Interests from Social Networks by Mining Users' Tagging Behaviour: A First Step Towards Adapting Social Information. *Data & Knowledge Engineering*, *108*, 15–29. Retrieved from http://10.0.3.248/j.datak.2016.12.003

Milgram, S. (1967). The Small-World Problem. *Psychology Today*. https://doi.org/10.1007/BF02717530

Our Story – Instagram. (n.d.). Retrieved February 12, 2018, from https://instagram-press.com/our-story/

Peng, R. (2015). The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance*, *12*(3), 30–32. Retrieved from http://10.0.4.87/j.1740-9713.2015.00827.x

Peng, R. D. (2009, July). Reproducible Research and Biostatistics. *Biostatistics*, pp. 405–408. Retrieved from http://10.0.4.69/biostatistics/kxp014

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=69861275&site=ehost-live

Pitic, A. G., Volovici, D., Tara, A., & Mite, A. C. (2013). Data Mining in Social Networks. *Revista Transilvania*, (2), 32–35. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=89005274&site=ehost-live

Platform Changelog - Instagram Developer Documentation. (n.d.). Retrieved February 14, 2018, from https://www.instagram.com/developer/changelog/

Postaddict.me. (2018). instagram-php-scraper. Retrieved from https://github.com/postaddictme/instagram-php-scraper

R: What is R? (n.d.). Retrieved February 13, 2018, from https://www.r-project.org/about.html

Richelson, J. T. (2015). *The US intelligence Community*. Westview Press.

Ronczkowski, M. R. (2011). *Terrorism and organized hate crime: Intelligence gathering, analysis and investigations*. CRC Press.

usairforce. (2018). U.S. Air Force (@usairforce) - Instagram Photos and Videos. Retrieved February 18, 2018, from https://www.instagram.com/p/BeQuT-8HfpC/?taken-by=usairforce

Ven, K. & Verelst, J. (2006). The Organizational Adoption of Open Source Server Software by Belgian Organizations. In E. Damiani, B. Fitzgerald, W. Scacchi, M. Scotto, & G. Succi (Eds.), *Open Source Systems* (pp. 111–122). Boston, MA: Springer US.

Wickham, H. (2015). *R Packages*. Sebastopol, CA: O'Reilly Media. Retrieved from http://r-pkgs.had.co.nz/

Wickham, H. & Grolemund, G. (2016). *R for Data Science*. Sebastopol, CA: O'Reilly Media. Retrieved from http://r4ds.had.co.nz/

Wickham, H. & Henry, L. (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.2. https://CRAN.R-project.org/package=tidyr

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
22-03-2018	Master's Thesis	October 2016 – March 2018

TITLE AND SUBTITLE

An Open Source Approach to Social Media Data Gathering

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kallhoff, Anthony J., Second Lieutenant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-18-M-130

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Intentionally Left Blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Modern usage of social media affords the military intelligence and analytic communities novel approaches to gather information. However, the tools and resources to develop these methodologies are still maturing. Furthermore, current data acquisition tools are not available to the DoD for all social media platforms. This thesis addresses a small subset of this problem by developing an open source methodological approach to collect and manage data from a popular social media site that has previously been inaccessible to defense intelligence organizations. This approach was operationalized via the R package called instaExtract, and an exemplar analysis was performed to demonstrate its application and efficiency for intelligence gathering.

15. SUBJECT TERMS
Data Mining, Social Media, Instagram, Data Acquisition, Data Tidying, Open-Source, R, GitHub, Web-based Hosting

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Bradley Boehmke, AFIT/ENS
U	**U**	**U**	**UU**	76	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636 (Bradley.Boehmke@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18